The Principles of Numerical Taxonomy applied to Multiple Choice Question Examinations Data.

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, largely it seems on the basis that the readership would not be able to understand it. That is for current readers to judge. 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

Numerical Taxonomy seeks structure amongst objects in a multi-dimensional measurement space. The response data to multiple choice examinations provides such a measurement space of high dimensionality, and the students/candidates are the objects being measured: the examiner is also an object in this space.
Many detailed methods have been described for use in numerical taxonomy. This paper discusses the special features of multiple choice examinations data, leading to choices of suitable taxonomic methods from amongst those available. The general form of interpretations and the possible value in taxonomic analysis of multiple choice examinations are reviewed.

Key words:

Numerical Taxonomy

Cluster Analysis

Systematics

Multiple Choice Questions (MCQ)

Introduction

One of the themes running through multiple choice question (MCQ) examinations research is efficiency: the reuse of the same material, and of the examination interaction data, for many purposes. Most obviously, examinations are used to grade the candidates by scoring (Buckley-Sharp & Harris (1971)). Grading provides a relative scaling, particularly a rank order, expressed via the scores. Viewed alternatively, MCQ examinations data may be used to assess the examination reliability, and to grade the individual questions. Standard methods are built into computer scoring systems (Buckley-Sharp & Harris (1972a)).

MCQ materials are used in closed examinations. They may also be used as quiz or self-testing aids; as discussion keys; or as aids to self-learning (Buckley-Sharp & Harris (1972b)). To aid in this reuse, MCQ may be stored in a recoverable form, and statistics on their effectiveness may be used to aid recall (Buckley-Sharp & Harris (1970)).

Here is outlined an altogether different use for MCQ examinations data: one which, as far as can be determined, has not been previously considered.

The MCQ Measurement Space

When students take an MCQ examination, they are each scored against the examiner. The examiner is assumed to be omniscient, although errors in the examiner's answers recorded for MCQ scoring are not unknown. Each score records the similarity between the respective student and the examiner, and a rank order of students is obtained.

The perfect student scores the maximum possible. There is only one way to obtain this score, and that is to submit a pattern of responses exactly matching that of the examiner.

(Given the examiner's finite probability of failing the omniscience test (qv), the perfect student may not score the maximum except by cheating!)

A hypothetical ignorant student obtains a score which can be characterised as random (Buckley-Sharp & Harris (1971)). Random scores would be obtained by tossing a coin to decide the response to each question in the examination. Of the 2^n patterns (where n=number of permitted responses) so obtainable, only one will be that of the perfect student. The majority, conventionally 95% of all possible patterns, will give scores which can be identified as random attempts. Yet it is inherent in the definition of a pattern which is random when compared to the examiner, that it is also random when compared to any other pattern.

The response data from an MCQ examination forms a multi-dimensional space: to an initial approximation it is n-dimensional (where n=number of permitted responses). The examiner is at one point in this space. Random scores form a uniform cloud in the measurement space so that most, say 95% of all possible patterns, may be recognised as being 'remote' from the examiner.

Real students, adequately tutored before an examination, do not show random response patterns when scored against the examiner. They are typically well inside the contour enclosing those 5% of patterns closest to the examiner. The question is - where exactly?

Traditional scoring is unidimensional: two students with the same score are deemed to have identical performance. Yet the measurement space is clearly multidimensional. Two students with very high scores must be similar because they are both close to a known point - the examiner. But two students with lower scores are not necessarily as close to each other as each is to the examiner.

Figure 1: Example responses by three students to three questions: and marginal scores.

Student	A	B	C	
Question				
1	correct	correct	error	2/3
2	correct	error	correct	2/3
3	error	correct	correct	2/3
	2/3	2/3	2/3	

In Figure 1, three students have made responses to three questions: the responses have already been interpreted against the examiner's pattern. While all three students attain the same score, as do all the questions, the response patterns leading to these scores are different. We do not know whether students as a whole locate uniformly in the measurement space surrounding the examiner, where the simple scores give the distances, but all directions are possible; or whether students cluster together into one or more discrete and separate locations.

Numerical Taxonomy - An Introduction

Numerical Taxonomy is the branch of statistics which evaluates the clustering of objects in measurement hyperspaces. Every problem analysis by this technique passes through four stages:

- Decide on the data parameters, and on the objects from which this data will be collected;
- Establish a suitable measurement of similarity and compute this value between every possible pair of objects;
- Decide on a suitable clustering algorithm, and use it to process the initial similarity matrix;
- Interpret the results.

The remainder of this paper discusses that range of data, objects, measurements and algorithms from within the canon of Numerical Taxonomy, which might reasonably be

applied to MCQ examinations. The general form of the likely interpretations will also be discussed. A subsequent paper (Buckley-Sharp, (1993)) demonstrates and evaluates an example of these techniques.

This discussion relies on the standard texts of Sokal & Sneath (1963), and Sneath & Sokal (1973). Both of these texts contain detailed bibliographies for further reading: this paper is not intended as a comprehensive review. Nevertheless, since Sokal and Sneath wrote their texts from backgrounds as biologists exploring the limitations of Linnean taxonomy, it is relevant here to test the details of numerical taxonomy separately against the particular requirements of MCQ examinations data. Vogt, Nagel & Sator (1987) provide a later review, written in the context of medical diagnosis and continuous variables: there are some additions in Vogt & Nagel (1992). Youngman (1979) also provides a short introduction with worked examples. Buckley-Sharp et al (1969) applied cluster analysis to MCQ examinations score and subscore data, but did not use response data directly for taxonomic study.

The Data and Objects

There is no shortage of data in an MCQ examination. It is provided by the responses from every student answering every question. The raw data is binary rather than continuous. The number of questions is generally large in comparison with the number of students, and particularly in comparison with the likely number of clusters. Thus, two difficulties commented upon in source texts - continuous or mixed data items, and small measurement spaces - are easily avoided.

The objects from whom the data are collected comprise all the students/candidates, **and including the examiner as an equal participant**. The statistical process treats the examiner as just another member of the set of objects. But it does become relevant to

draw upon the special properties of the examiner at the later interpretation stage, and this will be seen in the worked examples (Buckley-Sharp, 1993).

The available data is therefore taken from

Objects $\mathbf{O}_i$, i=1,m

Questions $\mathbf{Q}_p$, p=1,n

and comprises binary data

Responses $\mathbf{R}_{pi}$, p=1,n i=1,m

The use of the term 'Object' ($\mathbf{O}_i$) follows that of Vogt et al, in preference to the longer 'Operational Taxonomic Unit' (OTU) of Sokal & Sneath; but the concepts are identical.

Particularly for whole-year cohorts, and for major examinations, the number of response data items could easily range up to 100,000. Therefore issues of automated data collection, and of available facilities under particular computer hardware and software limitations become relevant. Thus, the computer programs written to support the examples have been limited to 120 objects. Even with this restriction, the requirement is approximately to create 7000 similarity coefficients, each from perhaps 1000 comparisons; to make 25 million comparison checks on the coefficients; and to modify the 7000 coefficients as clustering proceeds: and that from relatively simple methodology.

It is not a prior requirement that the data for numerical taxonomy is from a set of uncorrelated parameters. The effect of correlation between the data, which here means the responses to the questions, is to reduce the true dimensionality of the measurement space. Dimensional reduction may arise either from redundancy in the MCQ examination material itself, or from the recording technology.

If one question is duplicated, then the redundancy reduces the dimensionality of the space. This also occurs if one question overtly supplies the answer to another, or if two or more

questions are mutually exclusive, or if a student can evaluate the answer to one question logically from material displayed elsewhere in the examination. These opportunities do however depend, perhaps variably, on at least some skill by the student in recognising the duplication. Therefore, this effect has the interesting property of reducing the measurement dimensionality only in the vicinity of the likely participants, and particularly for the participants close to the examiner's point in the space. Participants therefore tend to use textual redundancy to gravitate towards the examiner in a way which is not seen for random attempts.

Aside from textual redundancy, when MCQ are presented as a stem followed by true/false alternatives, then each alternative is a data item. Contrast that common arrangement with the older 1-from-5 format, where the number of independent text items must be only one-fifth that of the equivalent true/false format examination.

There is also likely to be redundancy at the data collection stage, as can be seen by considering the usual methods.

For 1-from-5 questions, it is usual to offer five response boxes even though the boxes are mutually exclusive and only one of them can be selected. When computing the similarity coefficient (qv), the only two possible outcomes from five comparisons are then either {5-same + 0-different}, or {3-same + 2-different}. There will also be only two possible outcomes, {1-same} or {1-different}, from only one comparison, if the data from a question is first reduced to a single value eg, from the domain 'A'-'E' according to the single selection made.

For multiple true-false questions, two common formats are 'select if true' (one box offered), and the use of two boxes respectively for 'select if true' and 'select if false' with the don't know option shown by leaving both boxes blank. Obviously the two-box format

collects twice as many physical data items, and the clear distinction between the one-box 'leave blank if false' and the two-box don't know, represents real information. The two-box format really provides a single ternary data item, and not a quaternary or two true binaries because the option of selecting both boxes is prohibited. Specifically for the examiner as a participant, the don't know option is prohibited throughout, and the two boxes are perfectly correlated for all questions. For scoring purposes, Buckley-Sharp & Harris (1971) showed that the real information content of the one-box format for student participants was nearly as high as the two-box format. However, it does not necessarily follow that the one-box format is as suitable in the richer dimensionality of numerical taxonomy procedures.

This review shows that the true dimensionality of the measurement space is almost certain to be less than that suggested by the raw data collection procedure. Depending on the format of the examination and the method of recording responses, the dimensionality could be little better than a half, or even worse than one fifth of the total collected. Even so, MCQ examinations achieve value pre-eminently through quantity of material. The statistical methods will operate satisfactorily with redundant data included, and the later discussion on the interpretation will show that desirable cluster solutions are likely to be parsimonious rather than rich.

Therefore, it seems unnecessary to attempt any preprocessing of the raw data so as to reduce internal redundancy. Data reduction certainly can not increase the total information content. In this work, the response data has been retained as original binary items (<select> and <not-select>) from every response box offered to all the participants.

Similarity / Distance

Evaluating a taxonomy using the data collected from the objects requires a measure which expresses the similarity between every pair of objects. Some available measures show increasing values as the comparison objects become more similar: they are similarity coefficients. Other measures increase value as the objects become less similar: they are distance coefficients. Much has been written about the suitability of particular coefficients, and there is some interaction with the choice of a clustering algorithm (qv). A limited number of these measures has been chosen for review here.

The chosen coefficient **S** must be evaluated for every possible pair of objects $\mathbf{O}_i$, $\mathbf{O}_j$, where

$\mathbf{S}_{ij} = f(\mathbf{O}_i, \mathbf{O}_j, n) = f((\mathbf{R}_{ip}, \mathbf{R}_{jp}, p=1,n), n) : i=1,m\ j=1,m$

The response data must be compared in parallel from the two objects, and the comparison is repeated for all possible pairs: 'f' defines the algorithm of the coefficient **S**, whose calculation may also depend upon n, the number of comparisons made.

A particular axiom is that the similarity/distance of two objects is the same when viewed from any direction ie,

$f(\mathbf{O}_i,\mathbf{O}_j,n) = f(\mathbf{O}_j,\mathbf{O}_i,n)$

Therefore it is not necessary to compute the full m-square matrix of coefficients, since the matrix is symmetric.

Figure 2: Terminology for summarising raw data comparisons: and marginal totals.

Object$_i$ ->	<select>	<not-select>	
Object$_j$			
<select>	A	B	(A+B)
<not-select>	C	D	(C+D)
	(A+C)	(B+D)	(A+B+C+D)=n

As each pair of individual binary responses is compared, there are four possible results shown by the cell labels in Figure 2. On completion of all the comparisons, *A* is the count of occasions when both objects simultaneously selected an item. Similarly, *B, C* and *D* are counts of occasions for the other three response comparison possibilities. *A* and *D* are numbers of matches in the data; *B* and *C* are numbers of non-matches.

Sokal & Sneath (1963) discuss two parallel groups of coefficients which respectively exclude or include a component for negative matches (*D* in Figure 2). Then, some coefficients weight matches or non-matches more heavily on principle. It would seem inappropriate to exclude negative matches as they still represent agreement between objects. Sokal & Sneath were working with biological data where absence of a feature from two objects is not necessarily a coincidence. Still, they do not recommend any of the coefficients which exclude negative matches. There is much educational literature condemning selective weighting on the grounds that it cancels out in any large sample. Our own unpublished data confirms that when the examiner weights questions, supposedly on merit, then the rank order of students is not changed. Retaining negative matches, and eschewing selective weighting greatly reduces the variety of coefficients for consideration.

As part of the evaluation of these coefficients, mean values have been assessed for the comparison of two random patterns using a computer simulation. To allow for practical variation in MCQ answering situations, a range of biased 'coin tossing' parameters was used for both patterns in the comparisons. The biases were set to provide a <select> probability in 10% steps from 10% to 100%: 50% representing no net bias. By varying the biases for both patterns, one hundred combinations were tested for each coefficient studied. An obvious division which arose was into those coefficients where the mean similarity for random patterns remained constant, as distinct from those coefficients where the mean similarity varied depending on the biases.

Simple Matching: $(A+D)/(A+B+C+D)$

range 0 to 1

This coefficient has much to commend it, and it is also related to Euclidian distance (qv). It is easy to compute since it is only necessary to add up the agreements and not *A* and *D* separately. If the programming language has an Exclusive-OR function, it is even easier to add up the differences and use the converse function ie, $1-(B+C)/(A+B+C+D)$. The mean for random matches depends on the number of selections made.

Ochiai: AD / SQR$\{(A+B)(C+D)(A+C)(B+D)\}$

range 0 to 1

The denominator is the same as Phi, but the numerator is different. The mean value is variable for random matches.

Hamann: $\{(A+D)-(B+C)\}/(A+B+C+D)$

range -1 to +1

This is merely 2x<Simple Matching>-1, and has no other advantage.

Phi: $(AD - BC)$/SQR$\{(A+B)(A+C)(C+D)(B+D)\}$

range -1 to +1

The denominator is the same as for Ochiai, and the mean value is zero for all random matches. Phi is familiar to educational researchers as it is commonly used in MCQ question analysis. Phi is convertible to chi-square: Cramers V is the absolute value of phi.

Yule: $(AD - BC) / (AD + BC)$

range -1 to +1

The numerator is the same as phi. The mean value is zero for all random matches. Otherwise, Yule is only an empirical suggestion, and is not a serious contender.

At this point it is instructive to examine the scoring formulae commonly used in MCQ examinations. It might be supposed that they are suitable coefficients, since they are after all used to express a matching to the examiner.

Scoring: $A/(A+C) - B/(B+D)$

where the examiner is $Object_i$ in Figure 2

range -1 to +1

First, it should be noted that this formula is quite correct for both the one-box answering format and the two-box answering format. The reason is that the two-box format has no redundancy from the examiner's point of view. Every text question is offered as two exactly converse data items: the examination merely appears to be twice as long, and with the response boxes exactly 50%:50% correct:error. This coefficient has an advantage for scoring interpretation because it directly represents the proportion correct minus the proportion error.

The standard scoring formula is unsuitable for clustering because its value changes if the two objects are transposed. To meet the requirement for identity under transposition, the formula of a similarity coefficient must use terms from Figure 2 symmetrically about the diagonal. All the coefficients chosen to show here, except for the Scoring formula, meet that condition.

Distance measures include the average signed distance, the average absolute distance, and the Euclidean distance which is the square root of the sum of squared distances. Perhaps the most useful is the average squared Euclidean distance because this always ranges from 0 to 1 for binary data. The concept of distance is a geometric one, and is explained by viewing the measurement space.

Consider two axes, perpendicular and representing two independent data item vectors. Each object may record 0 or 1 on each axis so that the coordinate values (0,0), (0,1), (1,0) and (1,1) are all possible and can be plotted against the axes. When two objects are plotted simultaneously, their distance apart is calculated by the usual Pythagorean rule and can only be 0, 1 or $2^{0.5}$. If two objects are plotted in three dimensions, then the maximum distance between them is $3^{0.5}$. In n dimensions, the maximum distance between two objects is $n^{0.5}$; the maximum squared distance is n; and by dividing by n, the maximum average squared distance is 1, for any value of n.

Since the squared distance, using binary axes, is the sum of the non-matches between the two objects, then using the terminology of Figure 2,

Average Squared Distance: $(B+C) / (A+B+C+D)$

range 0 to 1

and this is exactly 1 - <Simple Matching>

The proportion of binary non-matches is also known directly from information theory as the Hamming distance. The simple computation, the direct relation to probably the best similarity measure, and the obvious geometric and information content interpretations make the average squared Euclidean distance a very attractive choice.

The final choice of a suitable coefficient is then affected by the choice of clustering algorithm (qv). The algorithm must recompute modified coefficients during the clustering process, and a coefficient is preferred if its basis is retained throughout the stages of computation. The distance measures meet this requirement. The average squared distance has been used in the examples reported elsewhere.

Clustering Algorithms

Available clustering methods fall into four classes

- Associative: where the m objects start with 1 each in m clusters, and the algorithm makes repeated fusions to end with 1 cluster containing all objects;
- Divisive: where the m objects start together in 1 cluster, and the algorithm makes repeated separations ie, the reverse of associative;
- Arbitrary with Relocation: where the m objects are at first arbitrarily distributed into a limited number of clusters, hopefully a similar number to a likely solution, and the algorithm seeks optimal relocations to find a best fit;
- Factoring: where the matrix of similarity coefficients is submitted to a standard factor analysis; leading significant factors are retained and trailing insignificant factors are discarded; the factor solution is rotated to simplify the loadings of the m objects onto the factors and thus identify cluster membership.

Most work has been done with associative clustering, and the methods can be shown to form a family (Abel & Williams, 1985), where the members are distinguished by different methods of recomputing the similarity/distance matrix after each fusion. Amongst the many methods are:

- Nearest neighbour: fuses two clusters where the distance between two objects in different clusters is a minimum;
- Centroid: fuses two clusters where the distance between their centroids is a minimum;
- Incremental Sum of Squares (ISS); fuses two clusters where the increment in the within-clusters sum of squares is a minimum: also called Ward's method.

All these methods are monothetic. Once a cluster is formed, it is never resplit.

The ISS method has become popular, and has been used in the examples reported elsewhere (Buckley-Sharp, 1993). The criterion is much like that found in ANOVA methods. The within-clusters sum of squares represents a collective view of the sizes of the clusters at any one level of the solution. It is plausible to seek a minimum of these sizes as a current optimum. ISS requires a distance measure rather than a similarity coefficient.

As the clustering proceeds, clusters reach different sizes. However, it is not necessary to return to the original response data to compute properties of new clusters. Instead, the matrix of distance coefficients is quite simply recalculated to reflect each new state. If two clusters i and j are being fused to form cluster k, then the distances d_{*k} from new cluster k to every other current cluster eg, cluster h, can be found for the ISS method, and the following apply:

$\mathbf{O}_k = \mathbf{O}_i \cup \mathbf{O}_j$

$n_k = n_i + n_j$

$d_{hk} = \{ (n_h+n_i).d_{hi} + (n_h+n_j).d_{hj} - n_h.d_{ij} \} / \{ n_h+n_i+n_j \}$

where d_{hi} is the previous distance from h to i

d_{hj} is the previous distance from h to j

d_{ij} is the previous distance from i to j

n_h is the number of members in cluster h

n_i is the number of members in cluster i

n_j is the number of members in cluster j

d_{hk} is the new distance from h to k

Since the fusion of clusters i and j is chosen because the coefficient d_{ij} is the smallest in the distance matrix, it is desirable that all the new values in the set d_{hk} are larger than d_{ij}. Therefore the series of smallest values in the matrix as clustering proceeds should form a monotonic increasing function, and this requirement is met by the ISS method.

Interpretation - Discussion

It is well known that numerical taxonomy will extract a cluster structure from random data. Interpretation of solutions must be plausible and cautious.

A taxonomy asserts that some objects are related ie, close to each other, while being distant from other objects or classes of objects. Much of Linnean taxonomy was assembled without the use of numerical methods. The notion that gulls and gannets are related together, but are less related to goats, would not seem to require statistical support. Equally, it should not take much enquiry to distinguish the knowledge base of lawyers from that of surgeons. It is perhaps when the observed objects are all somewhat similar that a more rigorous numerical method is needed, both to make the distinctions and to improve certainty in the interpretation. What types of interpretation might be relevant or useful in educational research?

First, a classification identifies groups whose membership may be considered adequately homogeneous. The combined data on group members then gives a better characterisation of the group than does the data on any one member. This characterisation may lead to a descriptive shorthand titling of the group. Titling should be done with caution lest the title itself should lead to invalid presumptions. If handling data from say 100 objects, the identification of only a smaller number of groups, say 2-5, with larger memberships is likely to give better characterisation of the groups. Even so, the identification of any groups depends upon the use which might be made of this classification.

A classification which arose from one data collection on a set of objects (in this case students) would not be much use unless a matching classification also arose from a separate data collection. If every occasion gave a unique result, we could not be certain of

the classification, and would not know what action to take. When data is used to make a clinical diagnostic classification then value is established if the classification leads to understanding of the disease, or to the selection of more effective treatments, or to the better prediction of prognosis. These criteria seem transferable to education.

Stability

In seeking a taxonomy within a group of students, stability of the classification is desirable. Buckley-Sharp et al (1969) discussed the educational issues related to stability. They only analysed one examination and so could not test for stability. Since students' knowledge is expected to evolve, stability may be limited in time, and may also be limited within subject areas. Where stability exists it is more likely to be found from two occasions in one subject area and very close in time. The examples of numerical taxonomy reported elsewhere (Buckley-Sharp, 1993) were chosen to meet this condition, and good stability was found. However, the limits of stability, when relaxing either time or subject constraints, have not been explored.

Understanding

The usual performance data on the MCQ examination is always available to help understand any classification obtained. Understanding may be assisted by the scores, or perhaps from a question analysis. Traditional question analysis is done on the criterion of high/low total scores. Question analysis might be repeated but using the taxonomy as the criterion.

Selecting Treatment

Different medical schools may publish different mission statements, and no doubt seek to admit students accordingly. Within a school a common official curriculum is normal. Taxonomic analysis might suggest that a common curriculum is inappropriate. After an MCQ examination, it is not difficult to conceive that the lowest scorers may need

remedial teaching. Taxonomic analysis might suggest either the form of that teaching, or more usefully that different forms of teaching are indicated for subgroups of low scorers.

Prognosis

This is the most problematic criterion of value. Simple statements like "If you go on like this, you will fail" presuppose that no effective treatment can be provided (qv). Progress on relating a taxonomy to prognosis might require trials, splitting identified groups into controls and educational treatments: then seeing if either the group becomes heterogenous, or the treated group merges with another previously identified group.

A difficulty in medical education is that schools are not necessarily seeking to make a uniform product. Students also pursue an enormous variety of actual routes through the supposedly uniform curriculum, picking up unique sets of experiences. Perhaps, from the detail of a taxonomy, it would only be necessary to distinguish groups representing various adequate performances from groups representing various inadequate performances. Although it might then be argued that the simple MCQ score meets this requirement, a taxonomic analysis might still define preferred remedial treatments for different inadequate performance groups.

Other Applications

This discussion has been entirely of the possibility and uses of a taxonomy of the examination participants ie, the students and the examiner. An examination is always an interaction between these participants and the material eg, MCQ, which is why the participants' taxonomy may vary by time and subject.

Taxonomies of MCQ material have been published before, but perhaps assuming that certain materials have inherent constant properties. It would be entirely feasible to study operational taxonomies of MCQ by variations of the methods outlined here.

Acknowledgements

I am indebted to colleagues in the former Department of Research and Service in Education of the Middlesex Hospital Medical School for help during the long gestation of this paper. Also, to the librarians of the University of London Senate House for extensive literature searches, during which we could find no prior similar work on this topic.

References

Abel D.J., Williams W.T., (1985), A Re-Examination of Four Classificatory Fusion Strategies, Computer Journal, **28/4**, 439-443.

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., (1970), The Banking of Multiple Choice Questions, British Journal of Medical Education, **4**, 42-52.

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., Harris F.T.C., (1972), MINISCORE - A Computer Program for Scoring Multiple Choice Tests and its Relation to Self-Learning Assessment Modules

(SLAM), International Journal of Mathematical Education in Science and Technology, **3**, 367-373.

Buckley-Sharp M.D., (1993), The Classification of Students in Two Multiple Choice Question Examinations by Numerical Taxonomy, (in preparation).

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.

Vogt W., Nagel D., (1992), Cluster Analysis in Diagnosis, Clinical Chemistry, **38/2**, 182-198.

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