J. R. Statist. Soc. A (1981),
144, *Part* 4, *pp.* 419–461

Statistical Modelling of Data on Teaching Styles

By MURRAY AITKIN, DOROTHY ANDERSON and JOHN HINDE

Centre for Applied Statistics, University of Lancaster, UK

[Read before the ROYAL STATISTICAL SOCIETY on Wednesday, May 6th, 1981, the President Professor D. R. Cox in the Chair]

SUMMARY

This paper presents the detailed statistical modelling of an extensive body of educational research data on teaching styles and pupil performance. Clustering of teachers into distinct teaching styles is carried out using a latent class model, and comparison of these latent classes for differences in pupil achievement is examined using unbalanced variance component ("mixed") models. Differences among the classes are altered by the probabilistic clustering of the latent class model compared to the original findings of the Teaching Styles project, and the statistical significance of the differences is substantially reduced when allowance is made for the correlation among children taught by the same teacher.

Keywords: TEACHING STYLES; CLUSTER ANALYSIS; LATENT CLASS ANALYSIS; VARIANCE COMPONENTS; EM ALGORITHM

1. HISTORY

THE publication by Neville Bennett of the Teaching Styles study (Bennett, 1976, subsequently abbreviated to TS) was an important contribution to classroom research in the UK. His findings received widespread publicity, and his major conclusion, that the results suggested unequivocally that formal methods of teaching are associated with greater progress in the basic skills, caused considerable controversy.

The statistical and educational bases of the conclusions were subsequently criticized by Gray and Satterly (1976). Bennett and Entwistle (1976) replied to these criticisms, and the statistical issues remained unresolved, despite further discussion of the statistical aspects in an unpublished report by Satterly and Gray (1976).

In 1977, Aitkin and Bennett applied to the SSRC for support for a research project on the statistical evaluation of the analysis of change in classroom-based research, using the Teaching Styles data as an illustration. The application was approved after some delay, and a research assistant (Jane Hesketh) was appointed from March 1st, 1979. A non-technical paper based on the final report on this project (HR 5710) will appear in the *British Journal of Educational Psychology* (Aitkin, Bennett and Hesketh, 1981). The present paper gives a detailed discussion of the statistical modelling of the Teaching Styles data, and relates it to the previous analysis by Bennett (1976) and the subsequent published discussion. The analysis carried out under HR 5710 was substantially extended by Dorothy Anderson and John Hinde under the research programme grant HR 6132, and the description given here includes these extensions.

2. THE TEACHING STYLES STUDY

We give now a brief description of the Teaching Styles study, condensed from TS Chapter 2, where a full description can be found.

The aims of the TS study were, briefly, to assess whether different teaching styles resulted in different pupil progress, and whether different types of pupil performed better under different styles of teaching. This was done in a seven-stage research project:

(1) The terms "progressive" and "traditional" were broken down into their constituent

elements through a review of the relevant literature and interviews with primary school teachers. These elements were operationalized by questionnaire items.

(2) After the questionnaire had been designed and piloted, it was administered to a large and representative sample of teachers.

(3) Cluster analysis was used to create a typology of teaching styles by grouping together teachers who responded similarly to the questionnaire items.

(4) The typology was validated by independent ratings based on classroom observation and the perceptions of pupils.

(5) A representative sample of teachers was selected from each teaching style by choosing those teachers closest to the central profile of each type.

(6) The pupils of the sampled teachers were followed through one school year, pre-tested on entry using a wide range of cognitive and affective tests, and post-tested prior to exit.

(7) To assess the relationship between pupil personality and teaching type, a typology of pupils was created by cluster analysis of the personality tests.

(8) The pupil typology was validated by observing the classroom behaviour of a 10 per cent sample of pupils.

(9) Hypotheses were tested statistically.

In the subsequent discussion, we shall be particularly concerned with the cluster analyses of (3) and (7) and the statistical analyses of (9).

The questionnaire contained 28 items covering six major areas of classroom behaviour: classroom management and organization, teacher control and sanctions, curriculum content and planning, instructional strategies, motivational techniques, and assessment procedures. In subsequent analyses, here as in Bennett (1976), the 28 items were coded into 38 binary items. These are summarized in Table 1.

The questionnaire was sent to head teachers in all the 871 primary schools in Lancashire and Cumbria for distribution to the 1500 third- and fourth-year class teachers. The final response rate was 88 per cent, giving a sample of 1258 teachers. The responses of third- and fourth-year teachers were found to be very similar, and the cluster analysis was based on the 468 fourth-year teachers.

A principal component analysis of the 38 items was first carried out (described in Bennett and Jordan, 1975). The first factor explained only 11 per cent of the variance, and there were seven factors with eigenvalues larger than unity. A varimax rotation of the seven-component analysis was carried out, and 19 of the 38 items had large loadings on one or more factors. These 19 items were retained for a cluster analysis of the teachers, the measure of (dis)similarity between teachers i and i' being the Euclidean distance $D_{ii'}$ between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, defined by

$$D_{ii'}^2 = \sum_{l=1}^{38} (x_{li} - x_{li'})^2$$

in the notation of Section 3.3. The clustering method was agglomerative (fusion), using iterative relocation to maximize between-cluster relative to within-cluster variation. Solutions from 3 to 22 clusters were obtained, and the 12 cluster solution chosen since it gave the overall maximum. Of the 468 teachers, 78 were not close to any of the 12 cluster centroids, and were not classified into any cluster. The 12 clusters were roughly ordered from extremely formal to extremely informal, though all clusters apart from the two extremes used some progressive and some traditional teaching practices.

In stage five of the research project, 37 teachers were selected to represent seven of the 12 teacher clusters: clusters 1 and 2 were informal, 3, 4 and 7 were mixed, and 11 and 12 were formal. Twelve teachers were initially chosen to represent each overall style, six each from clusters 1, 2, 11 and 12, and four each from 3, 4 and 7, with one additional informal teacher. The teachers selected were in each case those whose profiles most closely matched the group profile of their cluster. In the reanalysis the questionnaire data from one mixed style teacher could not be identified, and this teacher had to be omitted.

The teachers administered the attainment tests under normal classroom conditions, and the research team administered the personality tests within 1 month of the pupils' entry into their new fourth-year classes.

The personality tests were used to cluster the pupils into eight pupil types, using tests measuring extraversion, neuroticism, contentiousness, self-evaluation, anxiety, motivation, associability and conformity. The clustering method was the same as that used for the teachers.

Analysis of covariance was used to test for differences among the three overall teaching styles, adjusting for the pre-test. Highly significant differences among styles were found on all three achievement tests of reading, mathematics and English. In all cases the formal style had a significantly greater adjusted mean than the informal style.

Further investigation of differences in style means by sex of the child and personality cluster of the child were undertaken and reported, but no formal analyses of variance or covariance were presented. There were some indications of pre-test by style or sex interactions, but again no formal analyses were presented.

Further discussion of these results appears in the appropriate sections below.

3. A STATISTICAL MODEL FOR CLUSTERING

3.1. *Cluster Analysis*

Bennett and Jordan (1975) sounded a note of caution about their use of cluster analysis: "This study has demonstrated the value of this relatively untried statistical technique in creating meaningful types of teaching style. Its utility cannot be denied, but there are still uncertainties about the most appropriate similarity coefficients to use with differing kinds of data . . ., and these uncertainties should be borne in mind in assessing the clusters reported in this study."

In the reanalysis in this paper we use *mixture* models to represent nonhomogeneous populations.

Clustering methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory. Though theoretical difficulties remain in deciding on the number of clusters (see Section 3.4), for a given number of clusters the assignment of individuals to clusters is based on standard likelihood ratio methods analogous to those used in discriminant analysis.

3.2. *The Latent Class Model*

In re-examining the existence of distinguishable teaching styles, we begin with the original 38 binary items from the teacher questionnaire. The probability model adopted is a mixture or *latent class* model.

Suppose there are, in fact, k latent (i.e. unobservable) classes or types of teaching style, characterized by different frequencies of use of different behaviours. Let the proportions of each teaching style in the population be $\lambda_1, \lambda_2, ..., \lambda_k$, with $\Sigma_j \lambda_j = 1$. Given that a teacher is in the jth latent class, the probability that his vector $\mathbf{X}$ of responses ($\mathbf{X} = (X_1, X_2, ..., X_{38})$) takes the value $\mathbf{x}$ (where each of $x_1, ..., x_{38}$ is 0 or 1) is $P(\mathbf{X} = \mathbf{x} \mid j, \boldsymbol{\theta}_j)$, depending on a vector of parameters $\boldsymbol{\theta}_j$, this vector being (possibly) different for each latent class. The unconditional probability of the response $\mathbf{x}$, when we do not know the latent class of the teacher, is

$$P(\mathbf{X} = \mathbf{x}) = \sum_{j=1}^{k} P(\mathbf{X} = \mathbf{x} \mid j, \boldsymbol{\theta}_j)\, P(\text{teacher in class } j)$$

$$= \sum_{j=1}^{k} \lambda_j P(\mathbf{X} = \mathbf{x} \mid j, \boldsymbol{\theta}_j).$$

To specify the model completely, we need to specify how the probability $P(\mathbf{X} \mid j, \boldsymbol{\theta}_j)$ depends on $\boldsymbol{\theta}_j$. We postulate that, given the latent class to which a teacher belongs, his responses on the 38 binary items are *independent*:

$$P(\mathbf{X} = \mathbf{x} \mid j, \boldsymbol{\theta}_j) = \prod_{l=1}^{38} P(X_l = x_l \mid j, \theta_{jl}).$$

The parameter vector $\boldsymbol{\theta}_j$ now consists simply of 38 components $\theta_{j1}, \ldots, \theta_{j38}$, the lth of which, θ_{jl}, is the probability that a teacher in the jth class gives a 1 response to the lth item. A 1 response on the lth item has no effect on the probability of a 1 response on any other item, for teachers in the jth class.

This assumption of *conditional independence* has been widely used in latent class modelling in sociology (see Lazarsfeld and Henry, 1968; Goodman, 1978), and is directly analogous to the assumption, in the factor analysis model, that observed variables are conditionally independent given the factors, that is, that the observed correlations between items are due to the clustered nature of the population, and that within a cluster, the items are independent. The conditional independence assumption is difficult to verify in the TS data. Some relevant evidence is considered in Section 3.8.

3.3 *Maximum Likelihood Estimation*

The latent class model may be fitted to the data by maximum likelihood using the *EM* algorithm (Dempster, Laird and Rubin, 1977). This powerful method is available for a wide range of "missing data" problems. In this instance, the basis of the algorithm is the recognition that if the "missing data" had been observed, simple sufficient statistics for the parameters would be used for straightforward *ML* estimation. On the other hand, if the parameters of the model were known, then the missing data in the sufficient statistics could be estimated by their conditional expectations given the observed data.

The *EM* algorithm alternates these two procedures. In the *E*-step, the current parameter estimates are used to estimate the conditional expectations of the sufficient statistics, given the observed data. Then in the *M*-step, new *ML* parameter estimates are obtained from the current (expected) sufficient statistics. This sequence of alternate steps guarantees convergence to a local maximum of the likelihood function. However, in mixture models (and other non-standard models) multiple maxima of the likelihood function may be found, depending on the starting values chosen for the parameter estimates (see Section 3.5). Details of the *EM* algorithm for the latent class model were given by Aitkin in the discussion of Bartholomew (1980) and are not reproduced here.

The parameter estimates for the two- and three-latent class models are shown in Table 1. The item number corresponds to that in TS, pp. 166–169, the number in parentheses next to the item number being the number of this item in Table 2 of Bennett and Jordan (1975, p. 24). For the two-class model, the response probabilities marked † show large differences between the classes, indicating systematic differences in behaviour on these items for teachers in the two latent classes. For the three-class model, the response probabilities for Classes 1 and 2 are very close to those for the corresponding classes in the two-class model (though in most cases more widely separated), and the response probabilities for Class 3 are mostly between those for classes 1 and 2, except for those items marked with an asterisk. Thus Class 3 is to some extent intermediate between Classes 1 and 2. (The column headed δ in Table 1 is the set of discriminant function coefficients for discriminating between Classes 1 and 2. See Section 3.7 for discussion.)

3.4. *Significance of Latent Class Model*

Before attempting to interpret these results, we need to consider their statistical significance. Since this clustering model, like any other, will produce clusters with homogeneous random

TABLE 1

Two- and three-latent class parameter estimates ($100 \times \hat{\theta}_{jl}$) *for teacher data*

		Two-class model			*Three-class model*		
	Item	*Class 1*	*Class 2*	δ	*Class 1*	*Class 2*	*Class 3*
1 (1)	Pupils have choice in where to sit	22	43	−0·99	20	44	33
2	Pupils sit in groups of three or more	60	87†	−1·49	54	88	79
3 (2)	Pupils allocated to seating by ability	35	23	0·59	36	22	30
4	Pupils stay in same seats for most of day	91	63†	1·78	91	52	89
5 (3)	Pupils not allowed freedom of movement in classroom	97	54†	3·32	100	53	74
6	Pupils not allowed to talk freely	89	48†	2·17	94	50	61
7	Pupils expected to ask permission to leave room	97	76†	3·33	96	69	95
8 (4)	Pupils expected to be quiet	82	42†	1·84	92	39	56
9	Monitors appointed for special jobs	85	67	1·02	90	70	69
10 (5)	Pupils taken out of school regularly	32	60	−1·16	33	70	35
11	Timetable used for organizing work	90	66†	1·54	95	62	77
12	Use own materials rather than textbooks	19	49	−1·41	20	56	26
13	Pupils expected to know tables by heart	92	76	1·29	97	80	75‡
14	Pupils asked to find own reference materials	29	37	−0·37	28	39	34
15 (6)	Pupils given homework regularly	35	22	0·65	45	29	12‡
16 (i) (7)	Teacher talks to whole class	71	44	1·14	73	37	62
(ii) (8)	Pupils work in groups on teacher tasks	29	42	−0·58	24	45	38
(iii) (9)	Pupils work in groups on work of own choice	15	46†	−1·57	13	59	20
(iv) (10)	Pupils work individually on teacher tasks	55	37	0·73	57	32	50
(v) (11)	Pupils work individually on work of own choice	28	50	0·94	29	60	26‡
17	Explore concepts in number work	18	55†	−1·72	14	62	34
18	Encourage fluency in written English even if inaccurate	87	94	−0·85	87	95	90
19 (12)	Pupils' work marked or graded	43	14†	1·54	50	16	20
20	Spelling and grammatical errors corrected	84	68	0·91	86	64	78
21 (13)	Stars given to pupils who produce best work	57	29	1·18	65	30	34
22 (14)	Arithmetic tests given at least once a week	59	38	0·85	68	43	35‡
23 (15)	Spelling tests given at least once a week	73	51	0·95	83	56	46‡

TABLE 1

		Two-class model			*Three-class model*		
	Item	*Class 1*	*Class 2*	$\boldsymbol{\delta}$	*Class 1*	*Class 2*	*Class 3*
24	End of term tests given	66	44	0·90	75	48	42‡
25	Many pupils who create discipline problems	09	09	0·00	07	01	18‡
26	Verbal reproof sufficient	97	95	0·94	98	99	91‡
27 (i)	Discipline—extra work given	70	53	0·73	69	49	67
(ii) (16)	Smack	65	42	0·30	64	33	63
(iii)	Withdrawal of privileges	86	77	0·61	85	74	85
(iv)	Send to head teacher	24	17	0·44	21	13	28‡
(v) (17)	Send out of room	19	15	0·28	15	08	27‡
28 (i) (18)	Emphasis on separate subject teaching	85	50†	1·73	87	43	73
(ii)	Emphasis on aesthetic subject teaching	55	63	−0·33	53	61	63‡
(iii) (19)	Emphasis on integrated subject teaching	22	65†	−1·89	21	75	33
λ	Estimated proportion of teachers in each class	0·538	0·462		0·366	0·312	0·322

† Indicates an item with large differences in response probability between Classes 1 and 2.
‡ Indicates an item on which Class 3 is extreme.
$\boldsymbol{\delta}$ is the vector of discriminant function coefficients (see Section 3.7).

data, we need convincing evidence of the statistical significance of the latent-class clusters. There are two sources for this evidence. A graphical test is presented in Section 3.8. First, we consider a formal test of significance.

The usual asymptotic χ^2 distribution for the likelihood ratio test statistic does not apply in mixture models, including the latent class model, because the parameter value specified by the null hypothesis falls on the boundary of the parameter space.

In a two-component mixture model, we may write

$$f(\mathbf{x}) = \lambda f(\mathbf{x}\,|\,\boldsymbol{\theta}_1)+(1-\lambda)\,f(\mathbf{x}\,|\,\boldsymbol{\theta}_2),$$

where $0 \leqslant \lambda \leqslant 1$. To test for the existence of two components, we test the hypothesis $\lambda = 0$, which is on the boundary of the parameter space. An alternative formulation is to test the hypothesis $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, that the two components are identical. This does not correspond to a point on the boundary of the parameter space, but if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$,

$$f(\mathbf{x}) = \lambda f(\mathbf{x}\,|\,\boldsymbol{\theta}_1)+(1-\lambda)\,f(\mathbf{x}\,|\,\boldsymbol{\theta}_1) = f(\mathbf{x}\,|\,\boldsymbol{\theta}_1)$$

regardless of the value of λ. Thus the likelihood function is flat in λ if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Thus in the two-component mixture model the distribution of $-2\log l$ under the null hypothesis of a single population is unknown. There has been little empirical study of the distribution. In the case of a mixture of multivariate normals with a common covariance matrix, Wolfe (1970) suggested, on the basis of a small simulation, that $-2\log l$ could be approximated by χ^2_{2k}, where k is the dimension of $\mathbf{X}$. Hartigan (1977) suggested that the asymptotic distribution should be between χ^2_k and χ^2_{k+1}.

A solution to this problem using a Bayes formulation is given by Aitkin and Rubin (1981). The extensive computations required are not complete, and will be reported elsewhere. We give instead a small simulation which is sufficient to test the hypothesis of a homogeneous population at the 5 per cent level.

Suppose we have simulated s values of $-2\log l$ under the null hypothesis, and have one additional value of $-2\log l$ from the real data. If the null hypothesis is true, then all $(s+1)$ values come from the null distribution, and all $(s+1)!$ permutations of these values are equally probable. In $s!$ of these permutations, the value of $-2\log l$ from the real data will be the largest of the $(s+1)$ values. Thus under the null hypothesis, the probability that the real-data value of $-2\log l$ is larger than all s simulation values is $1/(s+1)$ (see Hope, 1968).

The 19 values of $-2\log l$ are shown in Table 2(a), generated from a single population in which the 38 items were independent, with response probabilities equal to those estimated from the real data. It is possible that the true asymptotic distribution may depend on the parameter values under the null hypothesis, though evidence below from the two-class null hypothesis suggests that this is not the case.

TABLE 2(a)
Nineteen simulation values of $-2\log l$ *for* H_0*: one class,* H_1*: Two classes*

58·0	59·1	59·5	62·0	64·2	65·6	67·8	69·1	69·8	70·3
71·3	76·4	78·4	80·3	81·8	83·0	84·1	84·3	84·4	

TABLE 2(b)
Nineteen simulation values of $-2\log l$ *for* H_0*: two classes,* H_1*: three classes*

56·8	62·1	62·6	63·7	68·3	68·8	69·5	69·7	72·1	74·1
74·6	75·3	76·3	77·1	77·8	79·5	85·9	87·0	87·8	

The value of $-2\log l$ from the real data is 775·8. This value obviously does not come from the same distribution as the 19. Formally, the hypothesis is rejected at the 5 per cent level. To test the hypothesis at the 1 per cent level would require the simulation of 99 values, which we did not attempt because of the substantial computer time required.

The (alternative) hypothesis of three classes is tested similarly against the null hypothesis of two classes by generating 19 values of $-2\log l$ from the two-class model, in which the parameters are set equal to the sample estimates from the TS two-class model. The simulation values are shown in Table 2(b). The value of $-2\log l$ from the real data is 184·7. The hypothesis is again rejected at the 5 per cent level.

Some evidence about the asymptotic distribution may be obtained from these simulation values. Normal probability plots of the two sets of values look quite similar, with means and standard deviations 72·1, 73·1 and 9·3, 8·6 respectively. If these values are from the *same* asymptotic distribution, they may be pooled and plotted as a single sample. The superimposed normal distribution (with pooled mean 72·6 and standard deviation 8·8) fits well except in the tails, which appear shorter than those of the normal.

It is obvious that none of χ^2_{38}, χ^2_{39} or χ^2_{76} provides an adequate representation of the asymptotic distribution (χ^2_{76} would have mean 76 and standard deviation 12·3).

3.5. *Multiple Maxima of the Likelihood Function*

Latent class models with four, five, ..., eight classes were also fitted to the TS data. Table 3 shows the values of $-2\log l$ for the successive hypotheses of an increasing number of latent classes. Although the test statistics appear large compared with the critical values for the homogeneity and two-class null hypotheses, we did not use or interpret models with more than three classes, for the following reason.

TABLE 3
Likelihood ratio test statistics for the number of latent classes of teaching style

Null hypothesis	$-2 \log l$
One class (homogeneity)	775·8
Two classes	184·7
Three classes	173·8
Four classes	142·5
Five classes	126·0
Six classes	121·9
Seven classes	96·3

Mixture models are known to possess multiple (local) maxima of the likelihood function (Hartigan, 1975, pp. 113–114), so that there could exist two or more different sets of parameter estimates, and probabilistic assignments of teachers to latent classes, which gave nearly equally good representations of the original data. In such cases it seems meaningless to talk of a set of uniquely defined "styles", characterized by the sets of item reponse probabilities, for these might have quite different interpretations for the multiple local maxima.

With the two-latent-class model, there was a unique maximum of the likelihood function, but with three or more latent classes, multiple maxima appeared, depending on the set of initial assignments of teachers to classes used to start off the iterative algorithm for the maximum likelihood estimates. One of the reasons for such local maxima is the very large number of parameters. For each extra latent class, an additional set of 39 parameters has to be estimated. For three latent classes, there are already 116 parameters, almost one-fourth of the number of observations. For four latent classes, the number of parameters is one-third of the number of observations.

In the three-class model, the estimates given in Table 1 were obtained for seven different sets of initial assignments, but for an eighth set convergence occurred to the parameter estimates given in Table 4.

The estimates for Classes 1 and 2 are generally similar to those in Table 1, though there are some discrepancies. The estimates for Class 3 are generally quite different. The value of $-2 \log l$ compared to the two-class model was 165·7, which is significant at the 5 per cent level. The

TABLE 4
Parameter estimates for three-class model, local maximum

Item	1	2	3	4	5	6	7	8	9	10	11	12	13
Class 1	22	61	38	92	95	86	96	80	87	29	90	14	92
2	50	91	24	63	38	29	68	32	63	56	60	45	74
3	27	70	19	66	95	95	96	75	76	66	83	60	83
Item	14	15	16 (i)	(ii)	(iii)	(iv)	(v)	17	18	19	20	21	22
Class 1	30	34	71	27	16	56	25	17	87	43	82	55	58
2	34	18	42	38	46	40	51	55	95	10	63	27	29
3	38	35	54	50	35	34	49	49	93	28	83	43	62
Item	23	24	25	26	27 (i)	(ii)	(iii)	(iv)	(v)	28 (i)	(ii)	(iii)	$\hat{\lambda}$
Class 1	72	64	10	97	70	63	86	24	20	92	57	11	0·482
2	44	40	09	94	53	39	77	17	17	53	65	60	0·318
3	70	61	06	96	56	57	80	19	08	37	52	86	0·199

estimated proportion of Class 3 in the mixture—20 per cent—is substantially smaller than in Table 1. Further discussion is given in Section 3.9.

3.6. *Validation of Model*

As indicated in Section 3.2, the validation of the latent class model is not a simple matter for the TS data. There are two aspects of the model which need validation: the number of latent classes, and the assumption of conditional independence of the individual items within each latent class.

The number of latent classes is investigated by the likelihood ratio test described in Section 3.4, and by the informal graphical procedure described in Section 3.8. The conditional independence assumption is particularly difficult to assess, because we do not have the actual class membership of each teacher: if this were known, then an independence model could be fitted to the subset of teachers in each class, and the goodness-of-fit of the independence model assessed for each class (though difficulties remain with the sparsity of the table). Goodman (1978, Chapter 8) has used the likelihood ratio test for goodness-of-fit in the latent class model applied to contingency tables, but the difficulty in the TS data is sparsity: we have 468 observations in a 2^{38} contingency table—there are 468 cells with one observation, and $2^{38}-468$ with none!

Some evidence for goodness-of-fit of the latent class model can be obtained from the informal graphical procedure referred to above. First, we consider an important form of the conditional probability of class membership.

3.7. *Discriminant Function*

We have

$$P(\text{class } j \mid \mathbf{X}=\mathbf{x}) = \lambda_j P(\mathbf{X}=\mathbf{x} \mid j, \boldsymbol{\theta}_j) \Big/ \sum_{j=1}^{k} \lambda_j P(\mathbf{X}=\mathbf{x} \mid j, \boldsymbol{\theta}_j)$$

with

$$P(\mathbf{X}=\mathbf{x} \mid j, \boldsymbol{\theta}_j) = \prod_{l=1}^{38} P(X_l = x_l \mid j, \theta_{jl})$$

from the model of conditional independence. The probability on the right-hand side can be written

$$P(X_l = x_l \mid j, \theta_{jl}) = \theta_{jl}^{x_l}(1-\theta_{jl})^{1-x_l}.$$

Then

$$P(\text{class } j \mid \mathbf{X}=\mathbf{x}) = c\lambda_j \prod_{l=1}^{38} \theta_{jl}^{x_l}(1-\theta_{jl})^{1-x_l},$$

where c is a proportionality constant, and hence

$$\log P(\text{class } j \mid \mathbf{X}=\mathbf{x}) = \log c + \log \lambda_j + \sum_{l=1}^{38} \log(1-\theta_{jl}) + \sum_{l=1}^{38} x_l \psi_{jl},$$

where ψ_{jl} is the *log-odds*:

$$\psi_{jl} = \log[\theta_{jl}/(1-\theta_{jl})].$$

We may write shortly

$$\log P(j \mid \mathbf{X}=\mathbf{x}) = \phi_j + \boldsymbol{\psi}_j' \mathbf{x},$$

where

$$\phi_j = \log c + \log \lambda_j + \sum_l \log(1-\theta_{jl})$$

and $\boldsymbol{\psi}_j$ is the vector of log-odds values for class j. In comparing the probabilities of class membership for classes j and j', we have

$$\log[P(j \mid \mathbf{X}=\mathbf{x})/P(j' \mid \mathbf{X}=\mathbf{x})] = \phi_j - \phi_{j'} + (\boldsymbol{\psi}_j - \boldsymbol{\psi}_{j'})'\mathbf{x}.$$

Thus a comparison of the probabilities of class membership may be based on the calculation of a *discriminant function* $\boldsymbol{\delta}'_{jj'}\mathbf{x}$, where

$$\begin{aligned}\delta_{jj',l} &= \psi_{jl} - \psi_{j'l}\\ &= \log\left[\frac{\theta_{jl}(1-\theta_{j'l})}{(1-\theta_{jl})\,\theta_{j'l}}\right]\end{aligned}$$

is the *log-odds ratio* for the lth item, for the two classes j and j'. The size of the coefficient $\delta_{jj',l}$ reflects the importance of the lth variable in discriminating between classes j and k. If $\theta_{jl} = \theta_{j'l}$ for a particular l, then $\delta_{jj',l} = 0$ for this l, and the lth variable does not contribute to the discrimination between the jth and j'th classes, though it may contribute to the discrimination between other classes.

In Table 1, the discriminant function coefficients for the two-class model are listed under the heading δ following the two-class probability estimates.

The *linearity* of the above discriminant function is a consequence of the conditional independence model fitted. If the models fitted to the separate classes contain interactions on the log-linear scale, the discriminant function will contain cross-products of the item variables x_l. A general discussion of discrimination between *observed* classes using binary variables is given by Anderson (1972) and by Goldstein and Dillon (1978).

Whether linear or non-linear, the discriminant function has a very useful *scaling* property. In Section 3.12, continuous latent variable models are briefly mentioned. In these models, the style latent variable is assumed to have a given (e.g. normal) distribution over the population, and the value of the latent variable for a given person is of interest. The latent class model, though based on the weaker assumption of a two-point distribution of style, nevertheless provides a scaling of the items, and an ordering of the teachers along a continuum defined by the discriminant function.

In the TS data, three latent classes can be identified, and there is no single continuum of teaching style. The formal–informal "dimension" does not adequately describe the "mixed" teachers, who are not intermediate between the other two styles on the disciplinary and testing items. A continuous latent variable model would need at least two factors to represent the data adequately.

3.8. *Graphical Tests*

The fundamental role of the discriminant function suggests a simple graphical test for the existence of latent classes. In the two-class model, there is just one discriminant function $\boldsymbol{\delta}'_{12}\mathbf{x}$. If $\mathbf{X}$ has a simple multinomial distribution, and there are no real latent classes, then the distribution of $\boldsymbol{\delta}'_{12}\mathbf{x}$ would be approximately normal, with mean $\boldsymbol{\delta}'_{12}\boldsymbol{\mu}$ and variance $\boldsymbol{\delta}'_{12}\Sigma\boldsymbol{\delta}_{12}$, where $\boldsymbol{\mu}$ and Σ are the mean and covariance matrix of the 38-dimensional multinomial distribution. On the other hand, if $\mathbf{X}$ has a multinomial mixture distribution with k components in the mixture, as implied by the latent class model, then $\boldsymbol{\delta}'_{12}\mathbf{x}$ would have approximately a normal mixture distribution. Thus if we inspect the marginal distribution of $\boldsymbol{\delta}'_{12}\mathbf{x}$, multimodality or other pronounced non-normality would lead us to suspect that the latent class model is appropriate.

A difficulty here is that the discriminant function coefficient $\boldsymbol{\delta}_{12}$ is unknown, and has to be estimated from the data, so is then a random variable. This means that the distribution of $\hat{\boldsymbol{\delta}}'_{12}\mathbf{x}$ would in general not be approximately normal, or mixed normal.

We may avoid this difficulty by considering an *a priori* linear function of $\mathbf{X}$ which does not depend on the data. If we recode all the items so that 1 represents the "formal" and 0 the "informal" end of the range, then an obvious choice is the total score $T = \Sigma x_l^*$ on all recoded items x_l^*, for $l = 1, ..., 38$, which we may call the TOTAL FORMALITY SCORE. Then if there are no latent classes, but a single homogeneous population, T will be approximately normal with mean μ and variance σ^2, while if there are k latent classes, and the conditional independence model with parameters θ_{il}^* and λ_i holds, then T will be approximately distributed as a normal mixture with k components in proportions $\lambda_1, ..., \lambda_k$, the mean and variance of the jth component being given by

$$\mu_j = \sum_{l=1}^{38} \theta_{jl}^*, \quad \sigma_j^2 = \sum_{l=1}^{38} \theta_{jl}^*(1-\theta_{jl}^*).$$

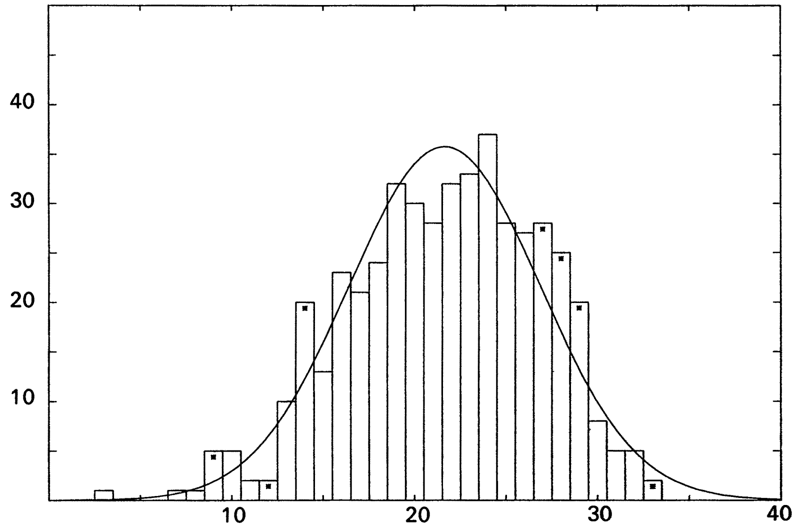
Fig. 1(a) shows the distribution of total formality score T for the 468 teachers, with the superimposed normal distribution with estimated mean $\hat{\mu} = 21{\cdot}64$ and variance $\hat{\sigma}^2 = 5{\cdot}22^2$. The overall goodness-of-fit χ^2 is 36·06, on 24 d.f., if the three smallest observations are grouped into a lowest class. Individual cell χ^2 values of more than 2·0 are indicated by asterisks on the appropriate cell. The fit is poor in both tails, and also in the right shoulder of the distribution.

Fig. 1(b) shows the superimposed mixture of two normals $N(25{\cdot}40, 2{\cdot}49^2)$ and $N(17{\cdot}29, 2{\cdot}79^2)$, in proportions 0·54 and 0·46 respectively. The fit in both the tails and the centre is bad, and the "dip" in the middle of the fitted distribution is not present in the data. The two-class distribution does not fit the data at all well.

Fig. 1(c) shows the superimposed mixture of three normals $N(26{\cdot}43, 2{\cdot}38^2)$, $N(21{\cdot}46, 2{\cdot}70^2)$ and $N(16{\cdot}18, 2{\cdot}75^2)$ in proportions 0·37, 0·32 and 0·31 respectively. The fit in the left tail is poor, but in the body of the distribution is good.

Since all three distributions fit badly in the left tail, we combine all the cells up to and including 12. The single normal then gives a goodness-of-fit χ^2 of 25·49, on 20 d.f., the two normals 53·90, and the three normals 18·17. The support for a three-component normal mixture, corresponding to a three-latent-class model, is quite strong.

Fig. 1(b) shows clearly that the apparently strong separation into two latent classes is misleading. The evidence strongly supports three overlapping, rather than two distinct, latent classes.



(a)

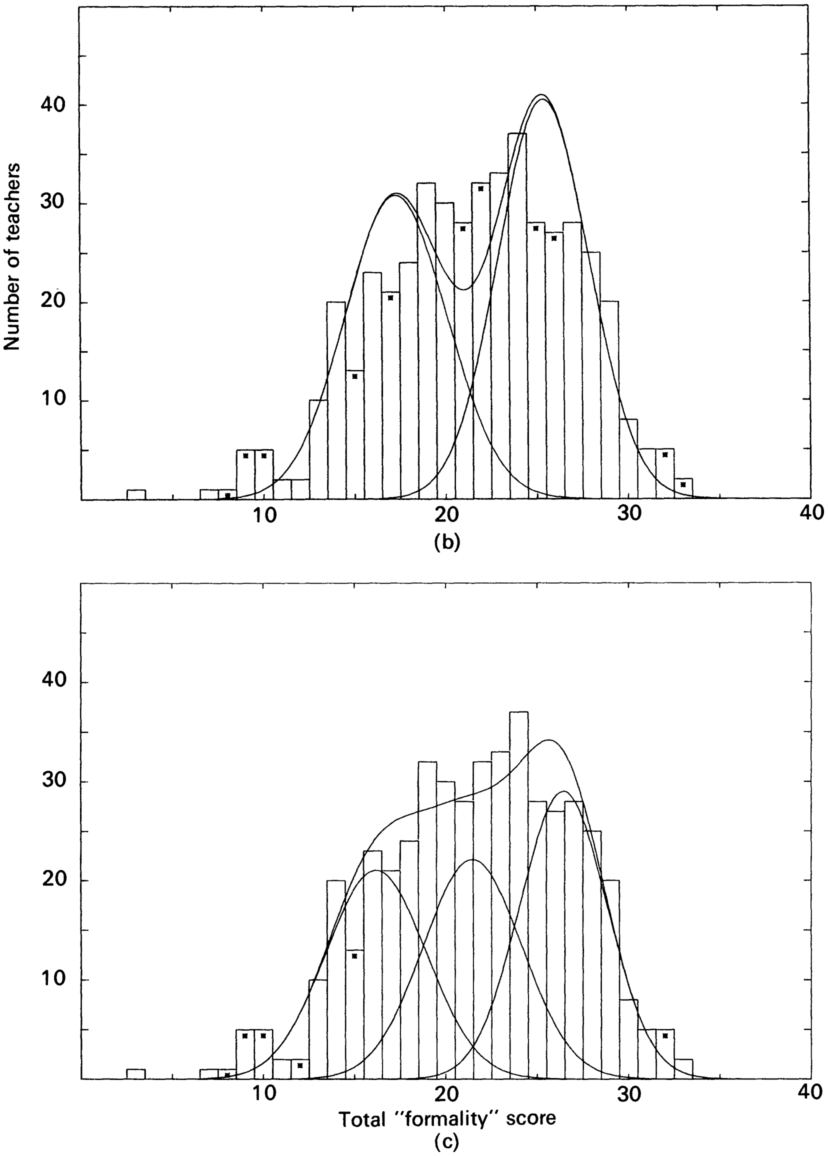


FIG. 1. (a) Homogeneous population. (b) Two classes. (c) Three classes.

3.9. *Interpretation of Classes*

We turn now to the interpretation of the teaching styles produced by the latent class model. A very clear and consistent pattern emerges in both the two- and three-class models. The first latent class is at the formal end of every item in the two-class model, and in the three class model except for items 27(iv) and (v), and 25. The second class is at the informal end of every item in the two-class model, and in the three class model apart from items 13, 15, 22, 23, 24 and 28(ii). Class 3 in the three-class model is intermediate between classes 1 and 2 on all items except those noted above.

Class 1 teachers almost all restrict children's movement and talking in the class room, while a large majority organize their work by timetable, emphasize separate subject teaching, and talk to the whole class, and a majority have pupils working individually on teacher tasks. Class 2 teachers are much less restrictive in their classroom organization, emphasize integrated subject teaching, and are likely to have pupils working individually or in groups on work of their own choice. Marking or grading of pupil's work is very uncommon in Class 2. The identification of Class 1 with a formal, and Class 2 with an informal, teaching style (as these terms were used in TS) is very clear.

Class 3 shares some of the characteristics of both the other classes. Like the formal teachers, their pupils stayed in the same seats for most of the day, were expected to ask permission to leave the room and were not taken out of school regularly, and the teachers used textbooks rather than their own materials, had similar teacher emphasis (Item 16) and similar disciplinary actions to the formal teachers. However, like the informal teachers, their pupils tended to sit in groups of three or more, they did not often mark or grade work, and did not give stars for good work. They placed greatest emphasis of all three classes on aesthetic subject teaching. It is notable that the Class 3 teachers were lowest in expecting pupils to know their tables by heart, in giving homework regularly, in giving weekly arithmetic or spelling tests, and end of term tests. Eighteen per cent of these teachers had many pupils who created discipline problems, compared with only 7 per cent of formal teachers and 1 per cent of informal teachers, and 9 per cent found a verbal reproof insufficient, compared with 2 per cent of formal and 1 per cent of informal teachers. Sending children out of the room, or to the head teacher, were more common disciplinary measures for Class 3 than for either of the other two classes.

While Class 3 shares some of the characteristics of each of the other two classes, and might therefore reasonably be called "mixed", the disciplinary problems and the low frequency of testing and assessment give this class a somewhat different character from that of the mixed style in TS.

We noted in Section 3.5 the occurrence of a local maximum of the likelihood function. The parameter estimates for Class 3 in Table 4 give a quite different interpretation of this class: these teachers were highest in restricting talking, in taking pupils out of school regularly, in using their own materials rather than text books, in asking pupils to find their own reference materials, in giving regular homework, in having pupils working in groups on teacher tasks, in correcting spelling and grammatical errors, in giving regular arithmetic tests and in emphasizing integratèd subject teaching. They were lowest in allocating pupils to seats by ability, in having pupils working individually on teacher tasks, in having many pupils with discipline problems, in sending children out of the room and in emphasizing separate subject teaching.

On other items, Class 3 were again intermediate between Class 1 and Class 2. It is tempting to conclude that the global maximum identifies Class 3 membership with an "uncertain" or "mixed-up" teaching style, and the local maximum identifies Class 3 membership with a "well-integrated" teaching style. Evidence for the first point will be brought out in Section 4.6. A further tentative conclusion is that teaching style is two-dimensional, and should be represented by continuous variables. This point is considered further in Section 3.12.

3.10. *Comparison with TS Clusters*

We consider now the comparison of the three latent classes described above with the 12 clusters described in TS. First we note a difficulty already referred to.

Since the result of probabilistic clustering is not an assignment to clusters but a set of posterior probabilities of class membership, it is not easy to present a simple table comparing the classes of teaching style for each clustering method. We present two tables. First, in Table 5 each teacher is formally assigned to the latent class to which he has the highest probability of belonging, and this assignment is compared with his membership in one of the twelve TS

TABLE 5
Latent class assignment and TS cluster membership for 468 teachers

Latent class	*TS cluster*													
	1	2	3	4	5	6	7	8	9	10	11	12	*Unclass*	*Total*
"Formal" (Class 1)	— (0)	— (0)	5 (21)	9 (27)	4 (15)	2 (5)	9 (30)	20 (67)	19 (53)	24 (77)	31 (79)	36 (100)	16 (21)	175
"Mixed" (Class 3)	1 (3)	13 (41)	11 (46)	2 (6)	7 (27)	26 (69)	19 (63)	6 (20)	14 (39)	6 (20)	8 (21)	— (0)	31 (39)	144
"Informal" (Class 2)	34 (97)	19 (59)	8 (33)	22 (67)	15 (58)	10 (26)	2 (7)	4 (13)	3 (8)	1 (3)	— (0)	— (0)	31 (40)	149
Total	35	32	24	33	26	38	30	30	36	31	39	36	78	468

The top entry is the number of teachers in each latent class who fall in the corresponding TS cluster, and the bottom entry is the percentage of teachers out of the total in this cluster.

clusters. It should be noted that 78 teachers were not assigned to any of the 12 TS clusters, as they were not close to any of the 12 cluster centroids. These teachers form the "unclassified" group in Table 5.

It is clear from Table 5 that only TS Clusters 1 and 12 correspond closely to the latent classes (2 and 1 respectively). About 40 per cent of TS Cluster 2 teachers are in latent Class 3, the "mixed" class, as are 20 per cent of TS Cluster 11 teachers. The remaining TS clusters are split across all three classes to varying degrees, the proportion of Class 1 teachers increasing, and of Class 2 teachers decreasing, fairly steadily from Cluster 1 to Cluster 12. Clusters 6 and 7 contain the greatest proportion of Class 3 teachers.

It was noted above that the formal assignment of teachers to latent classes overstates the information available from the probabilistic clustering. Since the conclusions drawn about pupil progress in Chapter 5 of TS depend critically on the cluster membership of the 37 teachers, we now consider the actual latent class membership for these teachers. Table 6 shows the

TABLE 6
Latent class probability and TS style category for 36 teachers

	TS style								
	Formal			*Mixed*			*Informal*		
Latent class:	1	3	2	1	3	2	1	3	2
	100	—	—	100	—	—	—	—	100
	100	—	—	100	—	—	—	—	100
	99	01	—	70	30	—	01	85	14
	99	01	—	12	88	—	—	—	100
	100	—	—	44	49	07	—	03	97
	100	—	—	01	98	01	—	—	100
	92	08	—	—	14	86	—	03	97
	100	—	—	100	—	—	—	—	100
	98	02	—	85	15	—	—	—	100
	100	—	—	11	89	—	—	—	100
	71	29	—	—	01	99	—	73	27
	94	06	—				—	36	64
							—	93	07

The entries are the probabilities of latent class membership (× 100) for the three-class model for 36 of the 37 teachers in TS Chapter 5.

probabilities of latent class membership for 36 of the teachers (one mixed TS style teacher could not be identified, and has been omitted from this table) for the three-class model.

The formal TS teachers, with one exception, have very high probabilities of belonging to Class 1. The one exception, in the three-class model, has a probability of 0·29 of belonging to Class 3, the "mixed" class. Nine of the 13 informal TS teachers in the three-class model have very high probabilities of belonging to Class 2, but three of the remaining four have high probabilities of belonging to Class 3, and the fourth is essentially unidentified. The mixed TS teachers are poorly identified: three clearly belong to Class 1, one to Class 2, and one to Class 3, while the remainder have substantial probabilities of belonging to two classes.

3.11. *Conclusion*

There is convincing statistical evidence, based on the latent class model, of three distinguishable teaching styles. Two of these correspond closely to the broad classes "formal" and "informal", as these terms were used in TS. The third class, called "mixed" here as in TS, is characterized by a low frequency of testing and assessment, and a relatively high frequency of disciplinary problems. The classification of the 36 teachers used in Chapter 5 of TS corresponds closely to the class membership probabilities for the three-class model for the formal teachers, less closely for the informal teachers, and poorly for the mixed teachers.

It is worth noting that the original cluster analysis used only 19 of the 38 binary items, these 19 having high loadings on one or more of the seven factors identified by the principal component analysis. Table 1 shows, however, that nearly all the 38 items discriminated among the three classes. The three-class model was fitted using only the 19 items, and the difference in $-2 \log l$ was 453·6 on 38 d.f. Substantial information about differences between the classes is lost if only 19 items are used.

3.12. *Continuous Latent Variable Models*

The latent class model of Section 1.2 assumes that there are discrete classes of teaching style. This model was developed and analysed because it corresponds closely to the original hypothesis in TS that there are distinguishable teaching styles.

However, an alternative model can be developed which regards teaching style as a continuum, with extremely formal styles at one end, and extremely informal styles at the other. All possible "degrees" of formality or informality might be possible, corresponding to intermediate points on this continuum.

A further possibility is that teaching style is multidimensional, and there is no single continuum of style.

Both of these possibilities can be modelled by replacing the discrete latent class model by a continuous latent variable model. Models of this kind are discussed by Bartholomew (1980), but were not used in the TS reanalysis, since maximum likelihood estimation in such models had not been successfully achieved at the time of the reanalysis. This has now been achieved by Bock and Aitkin (1981), and the TS data will be analysed by maximum likelihood using a two-factor model in a later paper. A two-factor model has already been fitted by Bartholomew (personal communication), using the moment method described in Bartholomew (1980).

4. THE RELATION OF TEACHING STYLE TO PUPIL PROGRESS

4.1. *Introduction*

In Chapter 5 of TS the relation between teaching style and pupil progress was investigated using an analysis of covariance model. The analysis was based on the individual pre-test and test scores of each child, the children being classified by the teaching style (formal, mixed, informal) of the teacher.

There has been considerable discussion, in the educational research literature, of the "unit of analysis" question: should the child or the classroom be treated as the "unit" on which statistical analysis is based? Gray and Satterly (1976) raised this question in their discussion on TS, and Satterly and Gray (1976) suggested the use of a mixed model in which teachers were treated as a random effect.

In this section we develop a "mixed" or variance component model for "clustered" or "nested" sample designs for the one-way analysis of covariance for pre-test/test situations. This model is then applied to the latent class membership for the 36 teachers described in Section 3.3.

4.2. *Variance Component Model for the One-way Classification*

Consider the following artificial experimental design. Thirty-six teachers are chosen randomly from a population of teachers, and at the beginnning of the school year are assigned randomly to classrooms. A random sample of 921 children from a large population is randomly divided into classes of about 25 children. Each teacher is randomly assigned to one of three teaching methods, and teaches one of the classes using the assigned method for the school year. At the beginning of the year the children are pre-tested, and at the end of the year are again tested on a standard achievement test. We want to determine whether the different teaching methods produce differences in the mean achievement score of children taught by each method.

This artificial design is a long way from that used in TS, or in most educational studies. It is introduced as the simplest model which allows unequivocal conclusions to be drawn about the effects of the experimental treatments, teaching methods here. The model is subsequently made successively more realistic.

Let Y_{pqr} denote the achievement test score, and x_{pqr} the pre-test score, of the rth child in the qth classroom, taught by method p, where $r = 1, ..., n_q$, $q = 1, ..., 36$, $p = 1, 2, 3$, $N = \Sigma_q n_q$. All subsequent analyses will be based on extensions of the following variance component model:

$$Y_{pqr} = \mu + \gamma x_{pqr} + \alpha_p + T_q + E_{pqr}.$$

Here T_q and E_{pqr} are mutually independent random variables, assumed to be normally distributed:

$$T_q \sim N(0, \sigma_T^2), \quad E_{pqr} \sim N(0, \sigma_E^2).$$

We may regard T_q as the "ability" of the qth teacher.

The α_p are constants with $\alpha_3 = 0$ (so that the model is of full rank—alternatively, we may take $\Sigma_p \alpha_p = 0$, as in Table 9), representing the mean achievement differences between methods 1 and 2, and method 3. The slope of the regression of test score Y on pre-test score is γ, assumed to be the same within each teaching method.

The T_q are treated as random variables rather than fixed constants because the teachers have been randomly selected from a population, and we are interested in modelling the variation in teacher ability in this population, as well as drawing inferences about the abilities of the particular teachers included in the sample. Teaching methods are represented by fixed constants because they are the unique set of experimental treatments under examination.

The properties of the above model are well known, and are described, for example, in Searle (1971, Chapters 9 and 10). A consequence of the random teacher effects is that the achievement scores of children within the same classroom are positively correlated:

$$\begin{aligned} \operatorname{var}(Y_{pqr}) &= \operatorname{var}(T_q + E_{pqr}) = \sigma_T^2 + \sigma_E^2, \\ \operatorname{cov}(Y_{pqr}, Y_{pqr'}) &= \operatorname{cov}(T_q + E_{pqr}, T_q + E_{pqr'}) = \operatorname{var}(T_q) = \sigma_T^2, \\ \operatorname{corr}(Y_{pqr}, Y_{pqr'}) &= \rho = \sigma_T^2/(\sigma_T^2 + \sigma_E^2). \end{aligned}$$

This intraclass correlation may be large if σ_T^2 is large compared with σ_E^2, and is zero only when $\sigma_T^2 = 0$, that is when there is *no* variation in ability among teachers in the teacher population, which will rarely happen in practice.

The above model may be extended to allow for pre-test by method interactions: it may happen that the slope of the regression of test on pre-test is different for different methods. A comparison of the methods then depends on the covariate value considered, and one method may be superior for low pre-test scores, while another is superior for high pre-test scores. The extended model is

$$Y_{pqr} = \mu+\gamma_p x_{pqr}+\alpha_p+T_q+E_{pqr},$$

and the regressions are now $\mu+\gamma_1 x_{1qr}+\alpha_1$ for method 1, $\mu+\gamma_2 x_{2qr}+\alpha_2$ for method 2 and $\mu+\gamma_3 x_{3qr}$ for method 3.

Unconditional conclusions about the relative superiority of one treatment to another are not possible in general with this extended model. Methods are available for drawing conditional conclusions, given the value of the pre-test score, based on the Johnson–Neyman "technique". We do not pursue them further here. The model may be further extended to include quadratic pre-test effects, sex of child, and interactions with sex.

In general, efficient (ML) estimation of the parameters in such models requires extensive iterative computation, even when the class sizes are equal. Details are given in Section 4.6.

The experimental question of interest is whether the different teaching methods affect the mean score of children in classrooms taught by each method. The null hypothesis of no difference among the methods is equivalent to $\alpha_1 = \alpha_2 = 0$. In the absence of covariates, and if the class sizes are all equal to n, this hypothesis may be tested from the following ANOVA table.

Source	*ss*	*d.f.*	*MS*	*EMS*
Among methods	SS_A	2	MS_A	$\lambda+n\sigma_T^2+\sigma_E^2$
Among teachers within methods	SS_B	33	MS_B	$n\sigma_T^2+\sigma_E^2$
Among teachers	SS_A+SS_A	35		
Within teachers	SS_E	$N-36$	MS_E	σ_E^2

When the class sizes are equal, the sums of squares SS_A, SS_B and SS_E are all independently distributed, SS_B and SS_E as multiples of χ^2 variables, the multipliers being the constants in the Expected Mean Square column, and SS_A as a multiple $(n\sigma_T^2+\sigma_E^2)$ of a non-central χ^2 variable, with non-centrality parameter λ which is zero when $\alpha_1 = \alpha_2 = 0$. The degrees of freedom for each χ^2 is given in the d.f. column. The test of the null hypothesis $\alpha_1 = \alpha_2 = 0$, which implies here that $\lambda = 0$, depends on whether $\sigma_T^2 = 0$ or not.

If $\sigma_T^2 \neq 0$, that is, there is positive correlation between the scores of children in the same classroom, the appropriate test is based on the ratio MS_A/MS_B, which under the null hypothesis has an $F_{2,33}$ distribution. This test is equivalent to a one-way ANOVA on the 36 class means over all children in each class—the class is the "unit of analysis".

However, if $\sigma_T^2 = 0$, then both SS_B and SS_E provide independent estimates of σ_E^2, and the appropriate test pools these two sums of squares, to give a test equivalent to a one-way ANOVA on all N children, since the scores of children in the same classroom are now independent—the child is the "unit of analysis".

To determine which test is appropriate, we examine the ratio MS_B/MS_E, which under the null hypothesis $\sigma_T^2 = 0$ has an $F_{33,N-36}$ distribution. Rejection of the null hypothesis leads us to conclude that $\sigma_T^2 \neq 0$, and the first test is appropriate. Failure to reject the null hypothesis, if n is reasonably large (and $n = 25$ certainly is), leads us to conclude that σ_T^2 is very small, or zero, and that the second test can be used.

Thus the "class versus child" decision may be based in this model on a preliminary test for the existence of positive intraclass correlation. This correlation may be estimated from the

ANOVA table: we have

$$\hat{\sigma}_E^2 = MS_E, \quad n\hat{\sigma}_T^2 + \hat{\sigma}_E^2 = MS_B,$$

so that

$$\hat{\rho} = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \hat{\sigma}_E^2} = \frac{MS_B - MS_E}{MS_B + (n-1)\,MS_E}.$$

It may happen that $\hat{\sigma}_T^2$ is negative; this is usually taken as evidence that σ_T^2 is zero or very close to zero.

4.3. *Unequal Class Sizes*

The above discussion is based on the assumptions that class sizes are all equal, and that there are no covariates. If this is not the case, the sums of squares SS_A and SS_B do not in general have multiples of χ^2 distributions, and the mean square ratios MS_A/MS_B and MS_B/MS_E do not have F-distributions. Further, the variance component estimators given above are not efficient. Efficient estimators of the variance components, and of the fixed effects, can be obtained by maximum likelihood in the unbalanced case, at the expense of considerable computation. Details are given in Section 4.6. The ANOVA table is replaced by an "analysis of deviance" table, in which successive values of $-2\log l$ from models of increasing complexity are differenced. The entries in the table have χ^2 distributions under the appropriate null hypotheses. For the simple one-way model with no covariates above, consistent estimates of the parameters may be obtained by treating teachers as a fixed effect, obtaining the usual SS breakdown for the hierarchical model, and then finding the expected values of the mean squares.

Using this method, an ANOVA table as set out above can be used for estimation and testing with the class size n replaced by a weighted average class size of

$$n^* = (N - \sum_q n_q^2/N)/35 = 25{\cdot}5$$

(Searle 1974, p. 474). The approximate F-tests for $\sigma_T^2 = 0$ and $\lambda = 0$ do not depend on the value of n^*, which affects only the variance component estimates. These F-tests are reported for comparison with the results in Bennett (1976). The likelihood ratio tests are also reported.

4.4. *The Effect of Non-random Assignment to Classes*

We began by considering an artificial fully randomized assignment of children to classes, and teachers to teaching methods. The reality of the classroom formation in TS is very different. First, teachers were not randomly assigned to methods: rather, teachers with existing styles were assigned (independently of the TS study) to intact classes. The greatest extent of randomness that could be hoped for is that the assignment of teachers was not based on the nature of pupils in the classes—that is, that teachers recognized as "formal" were not systematically assigned to classes which were below (or above) average on the pre-test.

If there *were* evidence of such an assignment bias, it would be difficult to draw general conclusions about differences in achievement between formal and informal teaching styles used on pupils of the *same* initial achievement, for teaching style and initial achievement would be at least partly confounded. Style differences, adjusted for initial achievement, would not necessarily correspond to those which would be found in a randomized experiment.

Since pupils were not randomly assigned to classes, we may expect that the 36 classes will differ systematically in their mean scores on the pre-test, such difference reflecting variation in the school populations, previous teachers and other systematic effects. The adjustment for the pre-test should then reduce the residual variation among teachers, and thus increase the sensitivity of the test for teaching style difference, since the variation among teaching styles would not be reduced by the pre-test adjustment if initial achievement and teaching style are not confounded.

Thus we may expect that the ANOVA variance component model, when applied to the TS study, will give interpretable results only if there are no systematic differences among teaching styles on the pre-test score. Even in this case, considerable care is needed in interpreting different styles as a *cause* of differential achievement. The data do not come from a randomized experiment, and there are many possible confounding variables. Discussions of such variables were given in TS, Bennett and Entwistle (1976) and Gray and Satterly (1976).

With these cautions in mind, we consider the results of the variance component models applied to the TS data in the next section.

A further difficulty, referred to several times previously, is that latent class membership is probabilistic, since class membership is not observable. An extended ANOVA model incorporating latent variables is necessary to model properly the full data: such a model is considered in Section 4.8. The analyses reported here are not based on a formal assignment of teachers to latent classes, but on the use of the probabilities of class membership as explanatory variables replacing the usual dummy variables. See Section 4.8 for details.

4.5. *Results for the TS Data*

We consider first the pre-test scores for reading, mathematics and English. Details of the tests are given in Chapter 5 of TS. All the test scores are normalized over reference populations. The one-way model of Section 4.2 is fitted to each of the pre-test scores using the consistent method for unequal class sizes. The ANOVA tables and style means are shown in Table 7, based on complete data for 921 children (although 950 children were analysed in TS, one complete classroom of 29 children was omitted in the reanalysis because the teacher's style could not be identified). The analysis of deviance tables are also given.

In all three cases, the F- and χ^2-values for style effects are very small, so there is no evidence of association of style with pre-test score.

This conclusion differs from that in TS. While there are some differences in the style pre-test means, due to the different style membership from the latent class model, the major difference

TABLE 7
ANOVA and analysis of deviance of pre-test scores

		Reading			*Mathematics*			*English*		
Source	*d.f.*	*SS*	*MS*	*Dev.*	*SS*	*MS*	*Dev.*	*SS*	*MS*	*Dev.*
Among styles	2	4495	2248	0·12	2224	1112	0·03	4197	2099	0·08
Among classrooms within styles	33	55980	1696		49334	1495		53880	1633	
Within classrooms	885	163540	184·8		106227	120·0		139293	157·4	

Variance component estimates:

	ANOVA	*ML*	*ANOVA*	*ML*	*ANOVA*	*ML*
$\hat{\sigma}_E^2$	184·8	184·9	120·0	120·0	157·4	157·5
$\hat{\sigma}_T^2$	59·3	64·2	53·9	57·5	57·9	63·3
$\hat{\rho}$	0·24	0·26	0·31	0·32	0·27	0·29

Estimated means	*ANOVA*	*ML*	*ANOVA*	*ML*	*ANOVA*	*ML*
Formal	101·8	98·4	100·5	98·8	103·4	101·2
Mixed	95·4	95·5	95·9	96·1	98·0	97·8
Informal	98·1	97·6	98·0	97·7	99·2	98·9

Dev., Deviance.

arises from the use of the residual variation among teachers as the error term in the ANOVA. It is clear from Table 7 that the use of the within-teacher variation (or the pooling of among- and within-teacher variation) would result in highly significant differences among styles on the pre-test.

The variance component estimates are also give in Table 7, based on both the approximate ANOVA and ML methods. The correlation between children's pre-test scores within classrooms is moderate, and certainly not zero.

We turn now to the test scores themselves. To give a direct comparison with the main results in TS Chapter 5, we present first the results of fitting two models, one with main effects of teaching styles and pre-test, and the other with an additional style by pre-test interaction (some evidence for the presence of such interactions was presented in TS Chapter 5). Table 8 gives the analysis of deviance for each test score, with the fitted regressions in Table 9.

TABLE 8
Analysis of deviance for test scores

		Deviance		
Source	*d.f.*	*Reading*	*Mathematics*	*English*
Style (adj. for pre-test)	2	0·69	2·70	5·12
Style × pre-test	2	0·34	1·10	4·01

TABLE 9
Parameter estimates from main effect models

	Style						
	F	*M*	*I*	*Pre-test* $(X-\bar{X})$	$\hat{\sigma}_T^2$	$\hat{\sigma}_E^2$	$\hat{\rho}$
Reading	105·0	103·6	106·0	0·76	23·7	43·2	0·35
Mathematics	103·6	99·8	103·5	0·87	18·7	50·2	0·27
English	107·4	103·4	105·8	0·74	9·85	46·8	0·17

Under the appropriate null hypothesis (no interaction, or no style mean differences) the corresponding deviance is distributed asymptotically as χ^2 with 2 d.f. None of the pre-test by style interactions is significant at the 10 per cent level. The main effect of style is significant for English at the 10 per cent level. No other effects are significant.

To investigate possible sex differences and interactions, a full analysis of the sex × style × pre-test model was carried out for each test score. The analysis of deviance tables are shown in Table 10.

The only important effect is the quadratic pre-test: the regression of test on pre-test is curved. Parameter estimates from the final models are given in Table 11. They differ negligibly from those in Table 9 except for reading, where the quadratic interaction reduces the differences between the mixed style and the other two for low or high pre-test values. No standard errors are reported for the parameter estimates as these are not obtained from the EM algorithm used to estimate the parameters.

The direction of the differences above is not consistent with those reported in TS. The formal classrooms do best in English, the informal classrooms do best in reading, formal and informal classes are very similar in mathematics, and the mixed classrooms do worst on all tests. It should

TABLE 10
Full analyses of sex × style × pre-test model

Source	*d.f.*	*Reading*	*Mathematics*	*English*
Style (adj. for pre-test)	2	0·69	2·70	5·12
Sex	1	0·11	0·01	0·56
Style × pre-test	2	0·33	1·09	4·06
Sex × pre-test	1	1·49	2·87	1·74
Sex × style	2	1·56	0·54	0·34
Pre-test × sex × style	2	(Model too large for program)		
Pre-test^2 (adj. for pre-test)	1	4·61	4·98	13·08
Style	2	0·67	2·74	5.14
Sex	1	0·06	0·00	0·51
Sex × pre-test	2	0·34	0·31	2·45
Style × pre-test^2	2	8·19	2·38	0·41

TABLE 11
Parameter estimates from final models

	Style				$(X-\bar{X})^2$		
	F	*M*	*I*	*Pre-test* $(X-\bar{X})$	*F*	*M*	*I*
Reading	104·8	101·8	106·1	0·76	$8{\cdot}6\times10^{-4}$	$7{\cdot}1\times10^{-3}$	$-4{\cdot}6\times10^{-4}$
Mathematics	104·1	100·2	103·9	0·88	$-2{\cdot}4\times10^{-3}$		
English	108·4	104·3	106·7	0·75	$-2{\cdot}9\times10^{-3}$		

be emphasized that these differences, though of educational significance, are not statistically significant. The non-significance results from the allowance for the random variation among classrooms (or the correlation between children in the same class), while the different direction of the differences results from the change in class membership for many of the "mixed" TS teachers, resulting from the probabilistic assignment by the latent class model. Fig. 2 gives a graphical comparison of the TS results with those given in Table 9.

The random variation among teachers or classrooms plays an important role. It can be seen from Tables 8 and 9 that the differences among the latent classes in mathematics are of the same magnitude as those in English, but the significance of the differences is much less for mathematics because the random variation among teachers is much greater. This variation is, in fact, quite substantial. This can be seen by considering the difference in ability between two randomly selected teachers of the same style (which equals the difference in mean achievement score for their classes, if the pre-test means are equal). If T_1 and T_2 are independently $N(0, \sigma_T^2)$, then $|T_1-T_2|$ has the truncated $N(0, 2\sigma_T^2)$ distribution on $(0, \infty)$, and $E(|T_1-T_2|) = 2\sigma_T/\sqrt{\pi} = 1{\cdot}13\sigma_T$. Thus in English, the average superiority of the better teacher over the poorer is 3·6 points, while in mathematics it is 4·9 points, and in reading 5·5 points. The largest style differences are 4·1 points in English, 3·9 points in mathematics and 2·4 points in reading. Thus individual variations in teacher ability are much more important for pupil achievement than teaching style differences.

The abilities of individual teachers could be estimated by treating them as a fixed, instead of a random, effect. However, a better method of estimation uses the additional information in the "prior distribution" of ability. It is natural to consider the "posterior distribution" of T given Y as containing all the information about teacher ability, given the prior distribution and the data

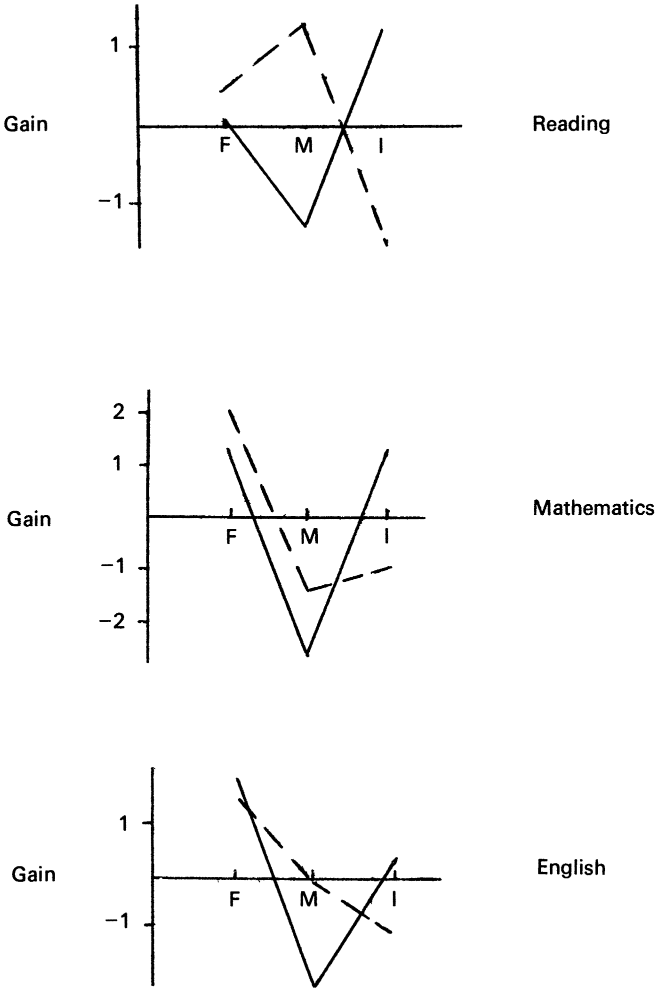


FIG. 2. Comparison of teaching style differences, TS and reanalysis. F, Formal; M, mixed; I, informal. - - -, TS; — reanalysis.

from pupils in each class. The *mean* of the posterior distribution is then the "expected" ability (see Section 4.8 for further details).

The posterior means for each teacher are shown in Table 12. There are several teachers—7, 9, 13, 17, 36—with consistently large means on all variables. Others show inconsistent high or low ability—2, 6, 12, 15, 19, 29. Of course it is possible that the effects being estimated contain other factors besides teacher ability. In any case, it is clear that very large variations can occur between classrooms which are not related to teaching style of the teacher.

4.6. *Maximum Likelihood Estimation in the Mixed Model*

Of the major statistical packages, only BMDP contains a general mixed model program (P3V), and this was not available at UMRCC at the time of the reanalysis. In the final report on HR 5710, several approximate analysis of variance methods were used, but these gave conflicting answers. A GENSTAT program was therefore developed for maximum likelihood estimation based on the *EM* algorithm. (The BMDP program is based on a combination of

TABLE 12
Posterior teacher ability means

Teacher	(*a*) *Reading*	(*b*) *Mathematics*	(*c*) *English*
1	0·1	−0·8	−1·0
2	5·1	−0·3	3·1
3	−2·0	−1·2	−2·7
4	−5·3	−3·0	−2·6
5	−1·0	−1·0	−0·7
6	6·0	−1·5	3·3
7	4·3	12·4	5·5
8	1·4	−0·7	−1·8
9	−9·8	−2·3	−3·5
10	1·0	−1·4	1·1
11	0·9	1·8	0·8
12	−7·5	2·5	−2·1
13	−7·1	−6·5	−3·4
14	−0·4	−2·5	2·8
15	−5·2	−5·0	−1·8
16	1·3	−2·7	0·1
17	−6·4	−7·2	−6·6
18	−2·0	−0·8	−1·4
19	4·8	1·5	3·2
20	11·9	4·6	−0·3
21	−3·3	−1·8	3·0
22	0·6	−1·0	−1·0
23	1·5	−2·4	−0·4
24	3·2	2·5	1·2
25	3·5	0·7	2·7
26	0·5	3·5	3·5
27	−1·9	1·0	2·7
28	1·7	0·1	−2·4
29	1·0	6·7	4·5
30	0·1	3·2	−1·6
31	0·1	−2·7	0·3
32	4·5	5·9	2·2
33	4·4	2·6	−2·3
34	−1·4	0·8	−0·7
35	3·7	3·0	1·6
36	−8·3	−7·8	−5·2

Fisher scoring and Newton–Raphson algorithms using second derivatives of the likelihood function.)

We may write the model in Section 4.2 as

$$\mathbf{Y} \mid \mathbf{T} \sim N_N(X\boldsymbol{\beta} + W\mathbf{T}, \sigma_E^2 I_N),$$
$$\mathbf{T} \sim N_Q(\mathbf{0}, \sigma_T^2 I_Q),$$

where $\mathbf{Y}$ is the N-vector of observations, $\boldsymbol{\beta}$ is the vector of regression coefficients of the "fixed effects" of dimension r, X is the $(N \times r)$ design matrix of the fixed effects, of rank r, $\mathbf{T}$ is the unobserved vector of abilities of the Q (= 36) teachers and $W$is the $N \times Q$ design matrix for $\mathbf{T}$.

The unconditional distribution of $\mathbf{Y}$ is multivariate normal with $E(\mathbf{Y}) = X\boldsymbol{\beta}$, $\mathscr{V}(\mathbf{Y}) = \sigma^2 H$, where

$$\sigma^2 = \sigma_E^2, \quad H = I + \gamma W W', \quad \gamma = \sigma_T^2/\sigma_E^2.$$

The ML estimates of $\boldsymbol{\beta}$, σ^2 and γ are found by differentiating the log-likelihood of $\mathbf{Y}$ in the usual manner. The likelihood equations, given by Hartley and Rao (1967), are not immediately soluble, and some iterative procedure is needed. Hemmerle and Hartley (1973) and Thompson (1975) give computational details for reducing the amount of work required to solve these equations.

The *EM* algorithm can be used to yield an iterative procedure, as described in Dempster *et al.* (1977). The "missing data" in this case are the teacher abilities **T**. If these had been observed, then the maximum likelihood estimates of $\boldsymbol{\beta}$, σ_E^2 and σ_T^2 would be

$$\left.\begin{aligned} \hat{\boldsymbol{\beta}} &= (X'X)^{-1}X'(\mathbf{Y}-W\mathbf{T}),\\ N\hat{\sigma}_N^2 &= (\mathbf{Y}-X\boldsymbol{\beta}-W\mathbf{T})'(\mathbf{Y}-X\hat{\boldsymbol{\beta}}-W\mathbf{T}),\\ &= (\mathbf{Y}-X\hat{\boldsymbol{\beta}})'(\mathbf{Y}-X\hat{\boldsymbol{\beta}})-2(\mathbf{Y}-X\hat{\boldsymbol{\beta}})'W\mathbf{T}+\mathbf{T}'W'W\mathbf{T},\\ Q\hat{\sigma}_T^2 &= \mathbf{T}'\mathbf{T}. \end{aligned}\right\} \qquad (1)$$

Thus the sufficient statistics involve the unknown **T** through **T**, **T′T** and $\mathbf{T}'W'W\mathbf{T}$, which are, in the *E*-step, replaced by their conditional expectations given the observed data **Y**. These are obtained from the conditional distribution of **T** given **Y**, which is

$$\mathbf{T}\,|\,\mathbf{Y} \sim N_Q(\gamma W'H^{-1}(\mathbf{Y}-X\boldsymbol{\beta}), \gamma\sigma^2(I_Q-\gamma W'H^{-1}W)).$$

Thus

$$\left.\begin{aligned} E(\mathbf{T}\,|\,\mathbf{Y}) &= \gamma W'H^{-1}(\mathbf{Y}-X\boldsymbol{\beta}),\\ E(\mathbf{T}'\mathbf{T}\,|\,\mathbf{Y}) &= E(\mathbf{T}'\,|\,\mathbf{Y})\,E(\mathbf{T}\,|\,\mathbf{Y}+\operatorname{tr}\mathscr{V}(\mathbf{T}\,|\,\mathbf{Y}))\\ &= \gamma^2(\mathbf{Y}-X\boldsymbol{\beta})'H^{-1}WW'H^{-1}(\mathbf{Y}-X\boldsymbol{\beta})+\gamma\sigma^2(Q-\gamma\operatorname{tr}(W'H^{-1}W)),\\ E(\mathbf{T}'W'W\mathbf{T}) &= \gamma^2(\mathbf{Y}-X\boldsymbol{\beta})'H^{-1}WW'WW'H^{-1}(\mathbf{Y}-X\boldsymbol{\beta})\\ &\quad+\gamma\sigma^2(N-\gamma\operatorname{tr}(W'WW'H^{-1}W)). \end{aligned}\right\} \qquad (2)$$

The algorithm begins with initial estimates of $\boldsymbol{\beta}$, σ_E^2 and σ_T^2, and substitutes these in (2). The conditional expectations are then resubstituted into (1) to give new parameter estimates, and this process continues until convergence. For the TS data, convergence occurs in 6–12 iterations, starting from $\gamma = 1$.

4.7. *Restricted Maximum Likelihood Estimation*

It is well known that the *ML* estimators of the variance components are biased. Patterson and Thompson (1971) derived unbiased estimators by restricting the estimation of the variance components to the error subspace orthogonal to the *ML* estimate of $\boldsymbol{\beta}$. The log-likelihood maximised is that of $S\mathbf{Y}$, where $S = I-X(X'X)^{-1}X'$. This estimation procedure is known as *restricted maximum likelihood* or *REML*, and in the absence of random effects it merely corrects the divisor of the estimator of σ^2. An advantage of *REML* estimation is that $\boldsymbol{\beta}$ is only calculated once. The *EM* algorithm for *REML* estimation is a slight variation on that described above, and is not discussed further.

A GENSTAT macro was developed for both *ML* and *REML* estimation, and was used for all the *ML* model fitting reported. All parameter estimates quoted are *REML* estimates. The *ML* estimates differ slightly, but lead to the same conclusions.

4.8. *Teacher Ability Estimates*

The conditional distribution of **T** given **Y** plays a fundamental role in the *EM* algorithm. The teacher effects are unobservable, so sufficient statistics involving them are replaced by their conditional expectations. If we view the $N(0, \sigma_T^2)$ distribution of teacher ability as a formal prior distribution, then the distribution of $\mathbf{T}\,|\,\mathbf{Y}$ is the posterior distribution, which contains all the information, from both prior and pupil data, about teacher ability. Since this posterior distribution is normal, its mean and (co)variance provide a complete summary of the ability distribution. For a *given* classroom, the posterior mean is the average ability of all teachers with the given pupil achievement.

The effect of incorporating the prior distribution of ability is to "shrink" differentially the ability estimates towards zero: small classes will produce greater shrinkage towards zero (the prior mean) than large classes. The degree of shrinkage depends on both n_q and γ, since

$$\hat{\mathbf{T}} = \hat{\gamma} W' \hat{H}^{-1}(\mathbf{Y} - X\hat{\boldsymbol{\beta}}) = \operatorname{diag}(\hat{\gamma}/(1+\hat{\gamma} n_q))\, W'(\mathbf{Y} - X\hat{\boldsymbol{\beta}})$$

and $W'(\mathbf{Y} - X\hat{\boldsymbol{\beta}})$ is simply the vector whose elements are the total for each classroom of the deviations of each pupil's score Y from the fitted value $X\hat{\beta}$. Let the classroom mean of these deviations for the q-th class be $\bar{d}_q$; then

$$\hat{T}_q = \hat{\gamma} n_q \bar{d}_q/(1+\hat{\gamma} n_q).$$

If $\hat{\gamma} n_q$ is large, then $\hat{T}_q \simeq \bar{d}_q$, and the teacher ability estimate is just the class mean deviation from the fitted regression. If $\hat{\gamma} n_q$ is small, then $\hat{T}_q$ is substantially shrunk towards zero. If γ is equal to zero then of course $\hat{T}_q = 0$, since there is then no variation in ability over teachers. Thus for reading, where $\hat{\gamma} = 23{\cdot}7/43{\cdot}4 = 0{\cdot}55$, the "shrinkage factor" $\hat{\gamma} n_q/(1+\hat{\gamma} n_q)$ is 0·86 for the smallest class of 11, but 0·95 for the largest class of 37. For English, where $\hat{\gamma} = 9{\cdot}9/46{\cdot}8 = 0{\cdot}21$, the corresponding values are 0·70 and 0·89.

4.9. *ML Estimation with the Latent Class Model*

The analyses reported above are all based on the variance component model in which the unobservable latent class dummy variables are replaced by their conditional expectations given the binary teacher response data, that is, by the posterior probabilities of class membership. The justification for this procedure is now given.

The full model for the binary behaviour variables and the pupil test data may be conveniently written by ordering the teachers so that the 36 included in the pupil study come first in the list of 468 teachers. Let $\mathbf{z}' = (z_1, z_2, z_3)$ be the vector of dummy variables indicating membership of the 468 teachers in latent Classes 1, 2 and 3 (with $\Sigma_{j=1}^{3} z_j = 1$). Then combining the models of Section 3.2 and Section 4.4, we have

$$\begin{aligned} P(\mathbf{X}_i = \mathbf{x}_i \mid z_{ji}) &= \sum_{l=1}^{38} P(X_{il} = x_{il} \mid j, \theta_{jl}), & i &= 1,\ldots,468,\\ Y_{ji'r} \mid z_{ji'}, T_{i'} &\sim N(\mu + \gamma u_{ji'r} + \alpha_j z_{ji'} + T_{i'}, \sigma_E^2), & i' &= 1,\ldots,36,\\ T_{i'} &\sim N(O, \sigma_T^2), & j &= 1,2,3,\\ P(z_{ji'} = 0) &= 1 - \lambda_j, & r &= 1,\ldots,n_i,\\ P(z_{ji'} = 1) &= \lambda_j \end{aligned}$$

with the T_i and z_{ji} being all independently distributed. Here u is used for the pre-test instead of x to avoid confusion with the behaviour variables.

In the latent class model we assumed that the binary behaviour variables X_{il} are conditionally independent given the latent class variables z_{ji}. We now extend this further, and assume that the X_{il} are also independent of the $Y_{jir'}$ conditional on the z_{ji}. This means that the binary behaviour variables tell us nothing about achievement which is not already contained in latent class membership.

Maximum likelihood estimation of all the parameters in the full model for all the data can be achieved using the *EM* algorithm with two stages of conditional expectations. First, suppose that the z_{ji} were observed. Then the parameters in the variance component model would be estimated using the *EM* algorithm as in Section 4.6, while the parameters in the latent class model would be estimated as in Section 3.3. Since T was not observed, the sufficient statistics involving **T** were replaced by their conditional expectations given **Y**. X was assumed to be known, but we now have part of X (the latent class dummies **z**) unobserved. Thus we need to take a further expectation, with respect to the conditional distribution of **z** given the observed data, of the *expected* sufficient statistics with respect to the conditional distribution of **T** given the observed data.

Since both **Y** and the binary behaviour variables **X** depend on **z**, the conditional distribution of **z** should be taken with respect to *both* **Y** and **X**. However, there is very little information in the class achievement data about the teaching style of the teacher relative to the information in the behaviour variables, and we may therefore take the conditional expectation of terms involving z with respect to the conditional distribution given **X** only.

It can be seen from equations (1) in Section 4.6 that the sufficient statistics depending on the design matrix X involve only linear and quadratic terms in z. These are therefore replaced by their conditional expectations given the behaviour variables X. Now

$$E(z_{ji} \mid X) = P(z_{ji} = 1 \mid X),$$
$$E(z_{ji}^2 \mid X) = E(z_{ji} \mid X)$$

since z_{ji} is either 0 or 1. Thus the linear terms in z are replaced by the posterior probabilities of class membership, but the quadratic terms (the diagonal terms in $X'X$ corresponding to the class membership dummies) are replaced, not by the squares of the probabilities, but by the probabilities themselves. This will slightly change the parameter estimates (both fixed effects and variance components) relative to those presented in Section 4.5, which are based on the replacement of the **z** by the posterior probabilities before the *EM* algorithm is applied to estimate the fixed effects and variance components. The extent of the change is unknown, though believed to be small, but will be investigated in a later paper.

4.10. *Conclusion*

The teaching style differences in achievement which were found in TS are not confirmed by the reanalysis. There are two reasons for this. First, the analysis of covariance model which includes the random effect of teachers results in greatly reduced significance of any differences, because of the large variation among teachers. Second, the clustering of teachers by the latent class model changes the nature of the differences among teaching styles.

It is of interest that the "mixed" style, which was characterized by a low frequency of testing and assessment, and a high frequency of disciplinary problems, shows consistently the poorest results in pupil performance.

5. PUPIL PERSONALITY

5.1. *Introduction*

In Chapter 8 of TS, pupils were clustered into eight personality types on the basis of eight personality variables. These personality types were then examined separately for teaching style differences, though a formal two-way ANCOVA was not used. The clustering was based on a Euclidean distance metric using iterative relocation, as for the clustering of teachers discussed in Section 2. We now describe a latent class model for clustering children when the multiple response variable are normally distributed.

5.2. *The Normal Mixture Model*

Suppose there are k latent classes of pupil personality, characterized by different mean values $\boldsymbol{\mu}_j$ of the vector of p personality variables, the covariance matrix $\boldsymbol{\Sigma}$ being common to all latent classes. Let the proportions of each personality type in the pupil population be λ_j, with $\Sigma_j \lambda_j = 1$. Given that a pupil is in the jth latent class, the probability distribution of his vector **X** of personality variables is assumed to be normal $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. The unconditional probability density function of **X**, when we do not know the latent class of the pupil, is

$$f(\mathbf{x}) = \sum_{j=1}^{k} \lambda_j f(\mathbf{x} \mid \boldsymbol{\mu}_j, \Sigma)$$

where

$$f(\mathbf{x} \mid \boldsymbol{\mu}_j, \Sigma) = (2\pi)^{-p/2} \mid \Sigma \mid^{-\frac{1}{2}} \exp - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)' \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_j).$$

An *EM* algorithm may be used to fit this model. Details were given by Day (1969) and Wolfe (1970); univariate versions are given by Aitkin and Tunnicliffe Wilson (1980).

The eight personality variables in TS were a subset of a larger original set of 15, and were the variables with high loadings on the factors in an earlier factor analysis of these 15 variables. In the reanalysis, we began again with the full set of 15 personality variables.

The normal mixture model was first fitted using 15 variables and 921 pupils, assuming a common but unspecified covariance matrix **Σ**. Convergence was extremely slow and huge amounts of time were required by the GENSTAT program without satisfactory convergence being achieved. Further, multiple maxima appeared even with two components.

A simplification of the model was then tried: the model was changed to *conditional independence*: instead of a common arbitrary covariance matrix, **Σ** was assumed to be diagonal. This model corresponds directly to the usual conditional independence factor model. The *EM* algorithm converged without difficulty, but again multiple maxima of the likelihood appeared. For the largest of these maxima, the reduction in $-2 \log L$ from one component (complete independence) to two components was 2517, and from two to three components was 772.

This model was also fitted using the eight TS personality variables. Multiple maxima again occurred even with two components, and the corresponding reductions were: 1–2, 656; 2–3, 164; 3–4, 187; 4–5, 110. A comparison of log-likelihood showed a substantial decrease in $-2 \log L$ from 8 to 15 variables: for the two-component case this was 861 (on 7 d.f.) and for the three-component case it was 608 (on 7 d.f.). Thus a substantial loss of information again occurs when the seven variables are omitted.

At this point the attempt to identify latent classes of pupils with different personalities was abandoned, since the multiple maxima meant that different definitions of such classes, and different sets of class membership probabilities were equally well supported by the data.

5.3. *Personality Factors*

An alternative model would allow continuous dimensions of personality rather than discrete classes. If the assumption of a latent class structure is replaced by a normally distributed variable or variables, the classical factor model is obtained:

$$\mathbf{X} \mid \mathbf{U} \sim N_p(\boldsymbol{\mu} + \Lambda\mathbf{U}, \Psi),$$
$$\mathbf{U} \sim N_r(0, \mathbf{I}),$$

where Λ is the (regression) matrix of factor loadings of the personality variables **X** on the factors **U**, and Ψ is the diagonal matrix of specific variances of the variables.

An important consequence of the factor model is that the vector of test scores **X** has marginally a multivariate normal distribution:

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Lambda\Lambda' + \Psi),$$

and, in particular, the individual test scores are normally distributed.

Efficient computational methods for fitting the factor model by maximum likelihood were given by Jöreskog (1967) and are available in several packages. We used the *EM* algorithm which is particularly simple to implement here, as noted by Dempster *et al.* (1977). Given the observed data $\mathscr{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$, the missing data are the factor scores $\mathscr{U} = (\mathbf{U}_1, \ldots, U_n)$. If $\mathscr{U}$ were observed, the sufficient statistics for Λ and Ψ would be S_{xu}, the usual corrected sums of squares and cross-products matrix, and S_{uu}, the *uncorrected* SP matrix for $\mathbf{U}$ (since $\mathbf{U}$ has zero mean). Then

$$\hat{\Lambda} = S_{uu}^{-1} S_{ux},$$
$$n\hat{\Psi} = \operatorname{diag}(S_{xx} - S_{xu} S_{uu}^{-1} S_{ux}),$$
$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}.$$

In the E-step, S_{uu} and S_{xu} are replaced by their conditional expectations given $\mathscr{X}$. Now

$$\mathbf{U} \mid \mathbf{X} \sim N_r(\Lambda'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}), \quad \mathbf{I}-\Lambda'\Sigma^{-1}\Lambda),$$

where

$$\Sigma = \Lambda\Lambda' + \Psi.$$

Thus

$$E(S_{uu} \mid \mathscr{X}) = n(\mathbf{I}-\Lambda'\Sigma^{-1}\Lambda) + \Lambda'\Sigma^{-1} S_{xx}, \boldsymbol{\Sigma}^{-1}\Lambda,$$
$$E(S_{xu} \mid \mathscr{X}) = S_{xx}\Sigma^{-1}\Lambda.$$

Alternate E- and M-steps lead to convergence to the unique maximum of the likelihood function. The algorithm is easily implemented in GENSTAT using initial estimates from a principal component analysis.

The restrictive feature of the factor model referred to above—marginal multivariate normality—is quite generally ignored. Indeed, the common view of factor analysis is that it is an exploratory data-analytic tool, and that therefore attention to distributional assumptions is unnecessary. For example, Taylor, in his chapter on factor analysis in O'Muircheartaigh and Payne (1977) argues: ". . . there is an attempt [in this chapter] to avoid embedding the description of the techniques in the framework of a multivariate normal distribution. Although this approach would allow a close link-up with classical statistical theory it would not be so readily applicable to practical data analysis. Most data are not multinormally distributed . . .".

But the likelihood ratio test for the number of factors, like any other statistical test for the structure of a multivariate normal covariance matrix, is critically affected by non-normality of the response variables. It is a common occurrence for many more factors to be found "significant" than can be interpreted. This is usually taken as evidence for the irrelevance of formal statistical tests. A more reasonable interpretation is that the distributional assumptions for the model are invalid.

The personality test scores provide a good example. The marginal distributions of the scores over the pupil sample show extreme skew, either positive or negative, on many variables. The classical factor model is therefore inappropriate without some substantial transformations of the scores.

A less restrictive version of the factor model might still be tenable. If $\mathbf{U}$ does not have a normal distribution, then the marginal distribution of $\mathbf{X}$ will be a normal mixture of some kind. It is possible to estimate the factor score distribution concurrently with the parameters of the factor model, using results due to Laird (1978). Details will be given elsewhere.

At this point we abandoned attempts to identify personality factors, and turned to the relation between achievement and the original personality variables.

5.4. *Regression of Achievement on Personality*

Preliminary regressions of the test scores on all 15 personality variables and pre-test score showed significant regressions for several personality variables for each of the test scores. The variance component model of Section 4.4 was therefore fitted with additional "covariates", these being the personality variables identified above. The models fitted were restricted to those using linear and quadratic pre-test, teaching style and the appropriate personality variables. Space constraints within the GENSTAT program prevented a full examination of style by personality interactions, though the interaction term for English was fitted and found to be very small, as were the individual interactions for mathematics. Results are shown in Table 13.

TABLE 13
Regression coefficients and deviance reductions for personality variables

			Regression coefficients		
Variable	*Range in TS data*	*Mean in TS data*	*Reading*	*Mathematics*	*English*
Psychoticism	(0–15)	3·3		−0·40	−0·20
Lie	(0–20)	11·0	−0·21	−0·15	
Introversion	(5–25)	14·5		−0·14	
Extroversion	(2–24)	18·6	0·19		
Contentiousness	(19–86)	43·7	−0·09		
Unsociability	(6–30)	12·9	0·14		
Conformity	(5–25)	18·9	0·19		
Deviance reduction (d.f.)			30·89 (5)	17·30 (3)	5·67 (1)

The deviance reduction is with respect to the model with style, pre-test and pre-test^2.

The effects of the personality variables are quite marked, especially for reading. The variables psychoticism, lie, introversion and unsociability were not included in the TS analysis. The first three of these, and contentiousness, are negatively associated with achievement in one or more test scores; extroversion and unsociability are positively associated with reading achievement.

6. GENERAL CONCLUSIONS ON STATISTICAL MODELLING

The analyses reported here were time-consuming, in both personal and computing terms, requiring the development of major GENSTAT, FORTRAN and GLIM programs, and very large amounts of computing time on the large UMRCC machines. Indeed, the Centre for Applied Statistics has become one of the major users of UMRCC time at Lancaster as a result of the Teaching Styles reanalysis.

The programs for clustering by the latent class model, and especially for fitting the variance component model, are of very general usefulness, though the latter is currently limited to one level of nesting. The GENSTAT macros can be used in any implementation of the system, though the version at Manchester has limited work space (88 000 real numbers) for large data sets.

The *EM* algorithm for maximum likelihood estimation with incomplete data is of outstanding importance. Though not the only possible computational method for such problems, it provides a unified theoretical approach, and the expected sufficient statistics computed at each step are often useful in themselves (the teacher ability estimates are an example). A major disadvantage is the lack of parameter standard errors, though it is formally possible to compute them from the conditional and unconditional cumulants, as described in Dempster *et al.* (1977). Since the likelihood in missing data models may be multimodal or otherwise badly behaved, considerable care should be taken in interpreting such standard errors when they are available.

The reanalysis of the clustering of the teachers, and the quite different results obtained, should sound a loud warning note to users of cluster analysis. It is a common practice to reduce dimensionality of questionnaire or other items by a factor or principal component analysis, and then use the reduced set of items, or the "important" principal variables themselves, as input to a cluster analysis program. The clusters produced in this way can only be regarded as arbitrary in the absence of any statistical model for the population, and "interpretability" of the clusters is a very poor substitute for statistical evidence of their reality.

It should be clear from the reanalysis that "battery reduction" methods should not be used *before* clustering by a mixture model: the discriminant function coefficients in the mixture model indicate the importance of the individual items, and items which discriminate effectively between latent classes need not be those showing large variability in the mixed population.

The general treatment of multi-stage sample designs requires variance component programs which can handle multiple random effects. There is a pressing need to develop such programs for large-scale surveys. Reports of users suggest that the existing BMDP program is both slow and very restricted in data size.

ACKNOWLEDGEMENTS

We are indebted to Darrell Bock for many discussions at an early stage of the preparation of this paper. We would also like to thank Harvey Goldstein, Ian Plewis and John Gray for helpful comments.

This research was supported by SSRC grants HR 5710 and HR 6132.

REFERENCES

AITKIN, M., BENNETT, S. N. and HESKETH, J. (1981). Teaching styles and pupil progress: a reanalysis. *Br. J. Educ. Psych.*, **51** (to appear).

AITKIN, M. and RUBIN, D. B. (1981). Estimation and hypothesis testing in finite mixture models. (In preparation.)

AITKIN, M. and TUNNICLIFFE WILSON, G. (1980). Mixture models, outliers and the EM algorithm. *Technometrics*, **22**, 325–331.

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, **59**, 19–35.

BARTHOLOMEW, D. J. (1980). Factor analysis for categorical data (with Discussion). *J. R. Statist. Soc.* B, **42**, 293–321.

BENNETT, N. (1976). *Teaching Styles and Pupil Progress.* London: Open Books.

BENNETT, N. and ENTWISTLE, N. (1976). Rite and wrong. A reply to "A Chapter of Errors". *Educ. Res.*, **19**, 217–222.

BENNETT, S. N. and JORDAN, J. (1975). A typology of teaching styles in primary schools. *Brit. J. Educ. Psychol.*, **45**, 20–28.

BOCK, R. D. and AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. (To appear in *Psychometrika.*)

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

GOLDSTEIN, M. and DILLON, W. R. (1978). *Discrete Discriminant Analysis.* New York: Wiley.

GOODMAN, L. A. (1978). *Analyzing Qualitative/Categorical Data.* London: Addison-Wesley.

GRAY, J. and SATTERLY, D. (1976). A chapter of errors. *Educ. Res.*, **19**, 45–56.

HARTIGAN, J. A. (1975). *Clustering Algorithms.* New York: Wiley.

—— (1977). Distribution problems in clustering. In *Classification and Clustering* (J. Van Ryzin, ed.). New York: Academic Press.

HARTLEY, H. O. and RAO, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93–108.

HEMMERLE, W. J. and HARTLEY, H. O. (1973). Computing maximum likelihood estimates for the mixed AOV model using the *W* transformation. *Technometrics*, **15**, 819–831.

HOPE, A. C. (1968). A simplified Monte Carlo significance test procedure. *J. R. Statist. Soc.* B, **30**, 582–598.

JÖRESKOG, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, **32**, 443–482.

LAIRD, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Ass.*, **73**, 805–811.

LAZARSFELD, P. F. and HENRY, N. W. (1968). *Latent Structure Analysis.* New York: Houghton Mifflin.

O'MUIRCHEARTAIGH, C. A. and PAYNE, C. (1977) (eds). *The Analysis of Survey Data*, Vols I and II. Chichester: Wiley.

PATTERSON, H. D. and THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

SATTERLY, D. and GRAY, J. (1976). Two statistical problems in classroom research. Unpublished.

SEARLE, S. R. (1971). *Linear Models.* New York: Wiley.

THOMPSON, R. (1975). A note on the *W* transformation. *Technometrics*, **17**, 511–512.

WOLFE, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multiv. Behav. Res.*, **5**, 329–350.

Discussion of the Paper by Professor Aitkin et al.

Mr D. Hutchison (National Children's Bureau, London): I would like to start by congratulating and thanking Professor Aitkin for his thorough and evenhanded investigation of what had become the vexed question of the soundness of the analyses in *Teaching Styles and Pupil Progress*, though this paper is of course more than a re-analysis.

I have already commented on earlier versions of this paper, so I do not intend to spend much time on technical aspects. However, I do feel that the authors seem to have missed an excellent opportunity for verifying their clustering model, by testing it on the large number of third year teachers, whose "response ... was found to be very similar to those of the fourth year teachers upon whom the analysis was based", instead of confining it to the fourth year, as they did.

I think that in general Professor Aitkin and his co-authors are correct to avoid the understandable tendency to claim that while the answers were wrong before, they have the right ones now. What they *do* have now is a method of clustering that is *statistically* soundly based given the assumptions they make. This is already an important achievement. It is necessary to emphasize this distinction since there are a number of reasons for not feeling that the definitive solution to the problems either of the grouping or of comparing teaching styles has been reached. These reasons in general have to do with the basis of methods of grouping. Firstly, it is of course slightly worrying that the method as developed by Professor Aitkin and his co-workers can only detect 3 clusters or groups in a sample of 468 teachers. Secondly, and rather more important, there is the question of what exactly constitutes a cluster. The word "cluster" to me carries with it some connotation of an entity with some internal cohesion. Is this so either in the data, or perhaps more relevant, in the staffroom? Or are we simply plotting a formality/informality dimension by locating points on it in much the same way as one draws a straight line by knowing two points on it? Possibly, as the writers have suggested, a two-dimensional solution would be more appropriate for the data, with a formality/informality dimension and perhaps a second dimension on the lines of an integrated style/disorganized style, as suggested by a synthesis of the global and local maxima described in this paper. Other speakers may well have more to say on this topic. The quasi-independence assumption, necessary as of course it is to obtaining *any* practicable solution, does seem somewhat unlikely in real life: if indeed formality/informality is an important distinction in the staffroom then surely teachers in either formal or informal groups will differ among themselves in the extent and enthusiasm with which they will adhere to the approach.

In statistical terms, there may well be an autocorrelation amongst answers in a group; or, looked at from another perspective, maybe a dimensional solution with an underlying vector would be more appropriate. Then there is the question of what the clusters themselves represent: is the major influence the teaching style, or is this simply acting as an alias for some other characteristic which is related to formality, but which has a greater effect on teaching success? Examples would be age or experience.

It is indeed excellent news that these data have now been deposited for general analysis with the SSRC Survey Archive. Though the deposition is somewhat belated, it still puts this study ahead of the majority studies in this country which bear upon education. I am inclined to feel that the most important findings of Professor Aitkin's reanalysis in educational terms are not the new grouping procedures, though these seem to be what have hit the headlines, or at least the specialist educational journals, but rather the finding that the original study used the wrong number of degrees of freedom, and that, when the correct number is used, none of the differences between groups are significant. Secondary analysis is obviously an important mechanism whereby other interested researchers may check for the possibility of such shortcomings. However, even secondary analysis cannot answer all the questions: a secondary analysis is only as good as the primarily-collected data. At this distance, and admittedly with hindsight, I find it hard to understand why anyone should have expected that style of teaching, as opposed to competence, conscientiousness, enthusiasm or experience, should make very much difference. Perhaps it was a popular belief which did not sort out sufficiently clearly the distinction between the *methods* of teaching and *competence* with which they were applied.

However, more important is the caveat that one can only get out of a cluster analysis (or any other sort of analysis) the type of variables one puts in, in the first place. Bennett's study was designed to compare Formal and Informal teaching or, to be more exact, teachers, and so other factors, which either corresponded to real-life grouping of teachers, or formed an important determinant of a 'style' of teaching or of pupil progress, might not have been collected in the first place.

Finally there is the question of the political impact of the original research compared with its statistical quality. The findings had considerable impact on educational policy. James Callaghan, the then Prime Minister, appeared to be influenced by Bennett's results in speeches on the "Great Debate" on education. According to Peter Wilby in *The Sunday Times*, Bennett's results helped swing primary schools back towards old-fashioned methods and away from experimentation. Yet the decision that these results were significant was based on an apparently simple mistake about the appropriate number of degrees of freedom to use.

On a wider perspective, nearly all major education research projects have been quite seriously academically criticized after their publication and after their initial public impact. The National Children's Bureau study of selective and non-selective schools and Rutter's study of primary schools, as well as Bennett's original book, are examples of this. If the reading public interested in education, and in this I include politicians and administrators as well as teachers and parents, are to become used to a pattern of publication of clear-cut results followed by their complete dismissal by some apparently equally eminent authority, then the credibility of any educational research and its statistical foundations will be, at least, very seriously eroded. How can we prevent this? It is tempting to remain in our ivory towers (or are they houses of cards?) and publish nothing except in specialist journals. Yet this is not acceptable. We are being paid public money to increase and improve the store of knowledge. If in our research we find out something we feel might be of importance to education (or any subject) we have a duty to make it available to all those who may be concerned.

So I shall make a few comments on how this could be handled. Obviously nothing can completely prevent malicious attacks by those with a vested and irrational belief in a point of view opposed to the results a project happens to find. However we can try to ensure that well-based expertise is included as far as possible. Some sponsors have insisted on an expert Advisory group to a project which they feel might be "sensitive" in its impact—DES as sponsor of the NCB selective and non-selective schools study is an example. Another, complementary, approach was that adopted by the SSRC on Professor Aitkin's project where the results had to be discussed at a seminar which might almost be called a "jury of peers" before the end of the project. No researcher can think of everything or be an expert in all types of technique. An advisory group during the life of the original project might have been able to offer advice on such questions as the size of the (teacher) sample, or possible third variables. A "jury of peers" could indicate possible alternative third variables or other objections, many of which could be readily dealt with, and also perhaps advise on the measure of confidence that the researcher and the public should repose in the results where the results are likely to be of public interest. While I cannot speak for Professor Aitkin, I feel that his project, excellent though it was in the first place, has been considerably improved as a result of this seminar. In passing, I should like to take this opportunity to comment on how impressed I am by the way this project has been developed since and, I think, as a result of, this seminar; and the large amount of work that must have been involved.

Finally I should like to extend my last remarks to express my considerable admiration at the impressive amount of work that this project has carried out during its life (and after) and to thank Professor Aitkin and his team for an extremely interesting paper, and one which is in my opinion one of the most important in this area to be published in this country for several years.

Professor D. J. Bartholomew (London School of Economics): The application of sophisticated statistical techniques to social data by statisticians is still something of a rarity. The authors are therefore to be congratulated on providing the Society with an opportunity to watch them at work and to discuss their methodology. Interest is sharpened by the obvious importance of the subject matter for educationalists and by the controversy which the original study stirred up a few years ago. I am personally grateful to the authors for making available the original data from the teachers' questionnaire, the re-analysis of which provides the basis for the following remarks. Although multivariate categorical data arise from almost every social survey it is remarkably difficult to acquire them in a form which permits secondary analyses. I hope that other investigators will follow the authors' example of making their data freely available so that teachers of statistics, and others, can help to realize their full potential.

The burden of my remarks concerns the methodological aspects and especially the model for clustering in Section 3. Bennett's original choice of cluster analysis following a principal component analysis on binary variables was clearly unsatisfactory. The authors were surely right to abandon this in favour of a properly structured probability model. It is not easy to understand, however, why they persisted with the notion that teachers could be classified into a small number of latent classes. There seem to be strong prior grounds for expecting human traits to vary continuously throughout the population and this points to a model with continuous latent variables.

The authors appear to recognize this, at least in retrospect, in Section 3.12 from which one deduces that they may have been unduly influenced by Bennett's original model. They also refer in that section to an analysis of my own using a latent variable model with two continuous latent dimensions. The results of this analysis throw an interesting light on the conclusions of this paper and will now be briefly outlined.

The model and method are as described in Bartholomew (1980). A two-factor model was fitted to the teaching styles data in the manner illustrated there on Upton's EEC data. A better method is to fit the two factors simultaneously but the program for this is not yet able to cope with 38 variables. I do not think that variations in the method of fitting would affect the conclusions now to be drawn. [The statements in the 1980 paper about the size of problem which could be handled were excessively pessimistic. There is no problem at all in fitting the one-factor model with 38 variables.] The first factor seems to correspond to the traditional/progressive dimension with positive loadings on most of those variables indicating a formal approach to teaching with negative loadings on the remainder. It is, however, the second factor which is of particular interest in the present context since it suggests that the questionnaire may have tapped another dimension which could be more relevant in interpreting the examination results. Excluding small loadings (arbitrarily defined) the larger negative loadings occur on variables 4, 7, 25, 27 (ii), 27 (iii), 27 (iv), 27 (v). Notice that most of these relate to discipline problems. The largest positive loadings occur on variables 26 (verbal reproof sufficient) and 15 (homework given regularly). There is a fairly long list of variables with moderate positive loadings: 3, 8, 10, 12, 13, 14, 16 (iii), 16 (v), 18, 19, 22, 23, 28 (iii). The striking thing about this list is that it includes a large number of variables which imply regularity in the teacher's behaviour—in giving homework, spelling and arithmetic tests and taking children out of school. Taken together these seem to present a picture of a teacher with a high degree of self-discipline which one might surmise would be conducive to discipline without formal sanctions in class. In other words, a high score on this dimension points to a well-organized person in control of the situation but also showing imagination and enterprise (e.g. variables 10, 18 and 28 (iii)). In short, this dimension seems to measure how "good" the teacher is at the job. One might expect this dimension to be more highly correlated with examination performances and I hope that such an analysis might be made.

The authors' "mixed-up" class is clearly picking up teachers who are extreme on this second dimension but by forcing the data into the 3-category mould is missing the real structure of the latent space.

I pointed out in the reply to the discussion of my 1980 paper that the 2-category latent class model emerged as a special case of the continuous latent variable model when the response function consisted of a single step. It is then easy to show that what I termed the *y*-scores are linearly related to the posterior probabilities of the latent class model and hence are functions of the discriminant scores. This is particularly interesting because I have been able to show that the *y*-scores are (almost) monotonic functions of $\Sigma \alpha_i x_i^*$ (where the α_i's are my factor loadings). In adopting the *total formality score* (Σx_j^*) the authors are using something which is highly correlated with the *y*-score. The distribution of the total formality scores given in the paper is much what I would have expected if the teachers were continuously distributed over the latent space. This makes me doubtful whether one can discriminate meaningfully between the 2- and 3-category latent class models on the basis of this distribution.

The authors strictures on the use of factor analysis in Section 5.3 are well made (though here, again, they seem to be attracted to a categorical latent space). The skewness of the marginal distributions should certainly not be overlooked. Notice, however, that factor analysis with ordered categories can come to the rescue. By categorizing the data the methods of Bartholomew (1980) become available without the need to make any distributional assumption. It may therefore be dubbed a "distribution free" method of factor analysis.

There are many other points which I would like to make on this stimulating paper but time calls a halt. The authors are deserving of a warm vote of thanks which I have pleasure in seconding.

The vote of thanks was passed by acclamation.

Dr G. W. Çran (Queen's University, Belfast): I should also like to thank Professor Aitkin and his co-authors for their interesting and wide-ranging paper. My main comments concern the use of the latent class model in cluster analysis.

The assumption of conditional independence (which the authors concede to be difficult to verify) implies that the response vectors within a latent class are sufficiently homogeneous so that within the class the binary responses may be considered independent. This assumption becomes more realistic as the number of latent classes increases, the number of binary responses remaining small. In the present study the number of latent classes is likely to be small and the number of binary responses (38) could be considered large. In addition Fig. 1 seems to suggest that homogeneity within two or three classes is unlikely.

Conditional independence also implies that $\log P(\mathbf{X} = \mathbf{x} \mid \text{class } j)$ is a linear function in $\mathbf{x}$. A simple extension to the latent class model would be the mixture model

$$P(\mathbf{X} = \mathbf{x}) = \sum_{j=1}^{k} \lambda_j P(\mathbf{X} = \mathbf{x} \mid \text{class } j),$$

where $\log P(\mathbf{X} = \mathbf{x} \mid \text{class } j)$ is a quadratic function in $\mathbf{x}$. Unfortunately the number of parameters to be estimated soon becomes excessively large as more interaction terms are introduced.

The following suggestion leads to a model containing a practicable number of parameters. The questionnaire contained 28 items, coded into 38 binary items, covering six major areas of classroom behaviour (these are listed in Section 2). For each area a score can be constructed based on those binary items pertaining to that area: hence

$$\mathbf{x} = (x_1, \cdots, x_{38}) \rightarrow \mathbf{y} = (y_1, \cdots, y_6).$$

The proposed model is a mixture model of the above type in terms of $\mathbf{y}$, with $\log P(\mathbf{Y} = \mathbf{y} \mid \text{class } j)$ a quadratic function of $\mathbf{y}$. This model has the advantages of a mixture model in cluster analysis as advocated by Professor Aitkin, uses all the information available on the major areas of classroom behaviour, and is more flexible than the conditional independence model.

I should also like to sound a note of caution on the conclusions drawn in Section 4. From Table 6 the 36 teachers can be allocated with reasonable confidence as follows: 17 in latent class 1, 11 in class 2 and 6 in class 3, with 2 unallocated. Hence the re-analysis does not provide much information on teachers belonging to class 3.

In conclusion, I would thank Professor Aitkin for coming to Belfast at this difficult time to present this stimulating paper.

Dr Fionn Murtagh (University College Dublin): I thank the authors for a very thorough paper. I would like to make two points regarding the cluster analysis.

(i) An important result which I have not seen is how the results of the cluster analysis used in the Teaching Styles study (i.e. the maximization of the between-group variance relative to the within-group variance) differed on being carried out using the full set of 38 items instead of the reduced set of 19 items. Without such a result, no judgement can be made on the suitability or otherwise of this method for the analysis of the type of data under consideration. The former results are certainly disputed; implicitly, the former method is also disputed; but I feel that the paper does not deal adequately with this former method. It seems, at any rate, that some other clustering method could usefully complement the latent class approach in dealing with an investigation of more than 3 classes. If I would like to stress a comparative clustering approach, it is because it is relatively all too easy to criticize the results of one clustering method by simply using another method.

(ii) The second point I will make concerns the sounding of a "loud warning note to users of cluster analysis": the overall conclusion (Section 6) warns strongly of unexpected effects in reducing an initial set of variables by means of a principal components or factor analysis. This conclusion is very interesting and demands further investigation. I think this could usefully be viewed as an aspect of the stability of the clustering method concerned, i.e. of the changes brought about in the clusters as a consequence of changes in the data. A recent comparative study of a number of clustering methods using simulated, error-perturbed data was that of Milligan (1980). A Monte Carlo study along these lines might be useful in establishing the extent of the problem indicated since, as was stated in the paper, the reduction of items or variables prior to cluster analysis is a common practice. So, in the case of principal components analysis, for instance, the "error" could be neatly characterized as the amount of variance ignored in using the reduced set of items.

Dr J. A. Wilson (Queen's University, Belfast): The reanalysis of Bennett's *Teaching Styles* data, by Professor Aitkin and his colleagues, has shown, firstly, that teacher ability is largely independent of teaching style, as re-defined by the latent class model, and secondly, that the three teaching styles so re-defined do not fit the uni-dimensional view of teaching to which Bennett's analysis subscribed.

In his preface to the *Teaching Styles* report Bennett had questioned the value of a dichotomous or uni-dimensional view of teaching. Despite this he proceeded to order his 12 teacher profiles, as defined by cluster analysis, on a formal–informal continuum before selecting and collapsing 7 of the 12 profiles into formal, mixed and informal categories. When his analysis showed that the more formal the teaching style the greater the degree of progress made by fourth-year primary school pupils it was hardly surprising that the evidence was seen to support the currently critical stance with regard to so-called progressive education. Bennett, in the report's conclusion, pointed out that "the findings will be disturbing .... since they indicate that the teaching approaches advocated by the Plowden Report .... often result in poorer academic performance". A TES headline of the time, "Bennett's bullets for Boyson's blunderbuss", conveys something of the educational partisanship evoked by the report.

When *Teaching Styles* appeared in 1976 several critics, but notably Gray and Satterly, drew attention to what they regarded as serious flaws in the analysis. Particular concern was expressed that an analysis based on pupils in intact classrooms should have failed to take account of variation between classrooms. There was also some concern at the way the cluster analysis data were used to form categories of formal, mixed and informal teachers. It was noted that while the 11+ examination and streaming were most prevalent in schools with formally taught classrooms, these confounding effects had been overlooked in the analysis of covariance.

Professor Aitkin's reanalysis has been particularly circumspect with regard to the unit of analysis problem; it has critically re-examined the teacher typology issue. It has given some useful answers and it has also raised one or two questions with methodological and educational implications.

Entwistle, in his foreword to Bennett's *Teaching Styles*, felt able to state that the findings suggested in unequivocal terms that formal methods of teaching were associated with greater progress in basic skills. In the light of that claim the single most important conclusion to come from the reanalysis is that "individual variations in teacher ability are more important than teaching style differences". What slight evidence there is of an association between teaching style and teacher ability, and none of this evidence is statistically significant, appears to indicate that the more successful teachers are those with well-defined teaching styles, whether formal or informal, and that the less successful are those "mixed-up" teachers who are low on assessment and high on problems of discipline.

The reanalysis has shown that, when teachers are assigned to discrete categories of teaching style, variation in teacher ability is greater within than between styles. The concept of teaching styles as discrete categories to which teachers, as types, may be assigned is analogous to a sociological rather than a psychological view of teaching. It does not permit of variation within the category and this may be a serious limitation on the data if, as the reanalysis indicates, the three styles do not locate on a single dimension, formal through mixed to informal, as Bennett's analysis had assumed. It is therefore of interest to note that a further analysis, in which the discrete latent class model is replaced by a continuous latent variable model, is to be reported.

Since the reanalysis, as now presented, has shown teacher ability, or effectiveness, to be largely independent of teaching style, it may be useful to reflect briefly on the ability data. These are presented (in Table 12 of Professor Aitkin's paper) as posterior teacher ability means for the basic skills of reading, mathematics and English for each of 36 teachers. The rank-order correlations between the sets of means range from 0·5 to 0·6, which leaves a considerable amount of unexplained variation as between teacher ability in these core areas. Furthermore, Gray (1979) has shown that within a particular area teachers are not consistently more effective. In a study of reading progress in 41 classrooms in two consecutive years he found that the correlation between residual gain scores by class in the first year and the second year was $r = 0{\cdot}01$. Some educational researchers have also become increasingly critical of the limitations of objective-type tests when these are used in studies of school or teacher effectiveness. They point to the care taken in the construction of such tests to minimize the effects of a taught curriculum.

May I say in conclusion that I am extremely grateful to Professor Aitkin and his colleagues for the very thorough work they have carried out on the *Teaching Styles* data, and to the Society for inviting me to contribute to this discussion.

Professor TONY GREENFIELD (The Queen's University of Belfast): In any study it is hard to be sure if the right variables have been observed. I asked my two teenage children, who have had more recent practical experience of observing teaching styles than most of us present, whether a formal style or an informal style was the more effective. They answered that far more important were whether or not a teacher was clearly confident about controlling and teaching the class and whether or not he could maintain a friendly manner. Perhaps in a future trial these points might be considered.

In Section 3.8 the authors introduce a total formality score. This is simply the sum of the scores of all 38 items. Would it not be more useful if weights were used according to the relative contributions of the different items? If so, how would the authors compute the best weights?

Miss Norma Reid (New University of Ulster): There is always someone who will get up during discussion of a paper and say: "I would not have started from here." I hope I may be forgiven for doing this by first saying how very interesting I have found this paper today, because I am working in a conceptually similar area, where I have encountered many of the difficulties discussed today. Specifically, I am looking at different pedagogies which exist in the training of student nurses in clinical areas, and then trying to relate different pedagogies to levels of attainment of student nurses. This is indeed a similar problem to the one being discussed today, and I have found that logistic discriminant analysis has a good deal to offer in this area. Logistic discriminant analysis is useful, for instance, because it is distribution free; it makes minimal assumptions about the data; it can handle data at any level of measurement; it can handle joint distributions of variables which are not necessarily independent.

In order to apply logistic discriminant analysis to this problem, it would be necessary, however, to have in addition to the 38 binary variables an independent method of classifying teachers into teaching styles. At the end of the day, any method of statistical analysis must produce latent classes which make sense to the educationalist—so why not start from here?—by building into the analysis the expertise of the educationalist. Indeed, it seems to me that the validation exercise reported on p. 48 of *Teaching Styles and Pupil Progress* does precisely this—so the data are already available. The case for any number of suspected latent groups can then be considered, by using the 38 binary items as input to a logistic discriminant analysis to see whether adequate discriminators can be found. Indeed, in addition to the 38 items, data on teacher ability or any other pertinent variable could also be input. This method would, I feel, make or break the case for the existence of latent groups.

If latent groups are indeed found to exist, the same method of analysis could be used to examine the relationship between teaching style and attainment scores. Using perhaps two groups of extreme scores, very high and very low, logistic discriminant analysis could be applied to a wider range of variables than are included in this paper. I would suspect, for instance, that social class of the pupil might well be strongly associated with attainment score. Also, I would be concerned about the effect of previous teaching styles to which the pupil has been exposed—what is the effect of a "formal" style in the third year followed by an "informal" style in the fourth year as compared to a consistent style in both years? [The Bennett study would appear to have pertinent data on this.] These are just two suggestions of extra variables which could be taken account of easily by logistic discriminant analysis. Teaching style, of course, would be an input. The logistic discriminant analysis would then tell us if teaching style is a significant factor, and if so, in conjunction with what other variables.

In conclusion, as a local group member, I would like to add my voice to those who have thanked not only our speaker Professor Aitkin for this fascinating paper, but also all of you who have travelled here today from London and Dublin. In the uncertain political climate prevailing here at the moment, all of us here greatly appreciate your support.

Mr M. R. Stevenson (Gallaher Ltd, Belfast): I would like to join in thanking Professor Aitkin for delivering a most interesting and useful paper, which may have a real effect on shaping the views of the consumer (i.e. parent) towards the various options available in primary education today. I should add that I personally am not involved in education, but come from a family with a history of employment in primary education, who would fall very much into the formal teaching class.

My comments therefore take the form of a subjective observation followed by a postulation.

Realizing that the study upon which the reanalysis is based was conducted in Lancashire and Cumbria I wonder how representative this would be of the primary education system in the U.K. This may not of course be absolutely critical to the final analysis, but observation of the primary schools in Northern Ireland does lead one to pose a question. In Northern Ireland, it is observed that there is a greater tendency for schools allocated to higher socio-economic areas to be more adventurous, i.e. teaching methods which would be classified informal, while those serving the lower socio-economic groups tend to be more traditional.

My postulation is this. Suppose inherent in the data are three classes as follows:

Class 1	Formal Teaching:	Lower socio-economic groups
Class 2	Informal Teaching:	Higher socio-economic groups
Class 3	Informal Teaching:	Lower socio-economic groups

I cannot of course speak for Lancashire or Cumbria, but in Northern Ireland there would almost certainly be problems with Class 3—particularly with respect to discipline, and hence a dilution in amount of informal freedom possible. I note that Class 3 as identified in Table 1 does at least match (and indeed is often more extreme than) Class 1 with regard to need for discipline, but perhaps to avoid straining an informal system to its limit where discipline is a problem, less homework and less class tests are noted than in Class 2.

I further note that while formal and informal classes give rise to approximately equal parameter estimates (Table 9), there is at least a hint of the "mixed" class giving lower parameter estimates. This would be consistent with what would arise under a hypothesis of lower scores both in lower socio-economic groups and by the informal teaching method. Class 1 scores are then reduced below their expected level because of the teaching method and become similar to Class 2, but Class 3 receiving both negative factors are below the aforementioned classes.

As I have already said, this is no more than a postulation, but I would be interested in hearing Professor Aitkin's comments on what confounding variables may be present in the study, and what action he would take to exclude them.

Dr A. R. Nicholson (Queen's University, Belfast): TS used the 19 (out of 38) best discriminating items as the composite criterion for clustering, and chose the 12 cluster solution. In the reanalysis all 38 items were used some of which may have decreased, rather than increased, cluster definition (introducing "noise", in fact). With a substantially new criterion it is perhaps not so surprising that the outcome is different.

Considering the 36/37 teachers chosen to represent

$$F\text{ormal}:M\text{ixed}:I\text{nformal} \equiv 12:12:13$$
$$\text{(one } M \text{ omitted in reanalysis: } 12:11:13).$$

In the re-analysis apparently we finish with

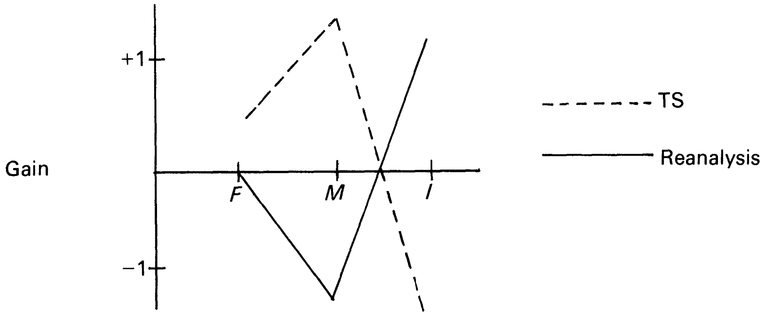
$$F':M':I' \equiv 17:8:11.$$

Considering the redistribution of the TS "Mixed" class

$$\begin{array}{rl} OUT & (11M \longrightarrow 5F', 4M', 2I') \\ IN & (12F \qquad\qquad\quad) \\ & (11M \longrightarrow 4M') \\ & (13I \longrightarrow 4M') \end{array}$$

it is not surprising that the results in Table 2 show a real change.

The new latent class (3) model reassigns teachers to the collapsed original categories: Formal, Mixed, Informal. This means that the summary Fig. 2, for reading (say), shows a distinct change due to the redesignation of whole classes of pupils:



In TS the teachers were representative of the categories F, M, I, are they still representative of the classes F', M', I'?

It is certainly right that published research should be subjected to scrutiny, etc. But what are we to make of educational research if we cannot know whether a reworking of the data by an alternative method of statistical analysis may not invalidate the conclusions?

For example

A. *The Heritability of IQ* by A. R. Jensen (based partly on work by C. Burt) challenged by L. J. Kamin on grounds of (a) honesty, (b) competence. Kamin's challenge deals largely with *interpretation* and does not directly question statistical procedures. (There is some selectivity in Kamin's attack.)

B. *The Stockholm Experiment* (Comprehensive Schools). Stockholm was divided into two sections: one to continue with the old selective schools system, the other to develop comprehensive schools. The study was reported by T. Husen and N. E. Svensson (1962). Part of the results indicated that gifted pupils do as well in a comprehensive system as in a selective system. In 1969, U. Dahllöf reviewed the whole study and found that Svensson's tests had not a high enough ceiling to evaluate adequately the learning done by the brighter pupils in the two sectors—invalidating this particular conclusion that had been drawn. This deals with *conceptual inadequacy*.

C. Today we have *Teaching Styles and Pupil Progress*: N. Bennett, reworked by Murray Aitkin (*et al.*).

Alternative decisions *38/19* items and *3/12* latent classes/clusters can give a significantly different outcome. It is good to establish this, but:

1. Is the whole reworking conceptually valid (as 1, 2, 3 above)?
2. Is cluster analysis sufficiently objective?
3. What does all this mean for educational research as a field for applied statistics?

Professor D. R. Cox (Department of Mathematics, Imperial College): This paper is a most impressive piece of work, in terms of the amount of detailed analysis accomplished, the broad technical statistical interest of the points raised and, above all, of the social importance of the topic. It is hard to comment on methodological issues without having looked at the data. Nevertheless it is tempting to ask whether some simpler analysis would not be effective, in particular in presenting the conclusions to a wider audience. For the part of the study connected with pupil attainment an approach would be

(i) to recognize that the unit of study is the class;
(ii) to reduce the set of scores for pupils in a class to suitable summary statistics;
(iii) to regress these summary statistics on explanatory variables, in particular on the subtotal questionnaire scores mentioned by Dr Cran.

Mr Antony Unwin (Trinity College, Dublin): I should like to raise two technical points and one general one. Although there has been justified criticism of the cluster analysis carried out by Bennett, he did attempt to identify outliers whereas Professor Aitkin and his co-authors have not done this. How should outliers be identified in a latent class model and would removing outliers from the Teaching Styles data affect the maximum likelihood estimates substantially?

Section 3.5 concerns multiple maxima of the likelihood function and the impression is given that any local maximum is a possible representation of the data. The existence of multiple maxima can bring computational difficulties in trying to find the global maximum but is otherwise irrelevant unless the likelihoods of the local maxima are close to that of the global maximum. From remarks Professor Aitkin made in presenting the paper it is clear that he is well aware of the problems associated with awkwardly shaped likelihood functions but I think there are more serious ones than that of the multiple local maxima referred to in the paper.

Like other speakers I am glad to hear that the Teaching Styles data have been made available for researchers. However, it would be a pity if statisticians got bogged down in reanalysing the same sets of data again and again as time series analysts often do. I fear we shall have many opportunities in the future of commenting on analyses of the Teaching Styles data at RSS meetings.

The following contributions were received in writing, after the meeting.

Mr Jeff Evans (Middlesex Polytechnic, Enfield): This reanalysis of the TS data is important, both because of the different substantive conclusions produced and because of its sobering message for users of statistical methodology. The demonstration here that the assignment of teachers to teaching styles depends on the choice of clustering techniques suggests that different variants of a specific statistical technique may tend to mould the data in distinctive ways. This points to a need for further investigation of several types: (i) routine examination of the sensitivity of research conclusions to the chosen variant of, say, factor analysis or clustering; (ii) further simulation studies of such sensitivity; (iii) reconsideration of the appropriateness of existing models for issues which are important in social and educational theory; for example, Pawson (1978) considers the limitations of regression models for representing social structural factors.

Another example comes from considering the notion of "teaching style", which is basic both to TS and to the reanalysis. Teaching style is clearly intended as a relatively permanent characteristic, independent of the context in which the behaviours are expressed; thus the θ_{jl} in Section 3.2 are independent of time and setting. This assumption contrasts sharply with the conclusions of many educational researchers that teachers' expressed views and behaviour may be context-specific (e.g. Keddie, 1971). Thus, for example, a teacher's response to survey items 10 ("pupils taken out of school regularly") and 12 ("use own materials rather than textbooks") might well depend on the resources available to the school, and those to items 16 (work with whole class, groups or individuals) and 20 ("spelling and grammatical errors corrected") might depend on the class size (s)he has to work with. Dealing with the context-specific nature of much human decision-making poses a substantial challenge for statistical modelling.

Mr B. S. Everitt (Institute of Psychiatry): I would first like to congratulate the authors on a paper notable for the manner in which it brings relatively sophisticated statistical techniques to bear on a problem of considerable importance. Of such importance in fact, that *The Sunday Times* itself was persuaded to mention, even if somewhat grudgingly, latent class analysis and unbalanced variance component models! My own comments are restricted to two aspects of the paper, namely determining the number of classes in the latent class model, and the use of the normal mixtures model.

That the usual asymptotic χ^2 distribution for the likelihood ratio test statistic does not apply to mixture models has, as the authors mention, been commented on by a number of workers including Wolfe, Hartigan and also Binder in his Ph.D. thesis on cluster analysis. Therefore it is of interest to use simulation methods to obtain evidence about the possible asymptotic distribution, as is done in Section 3.4 of this paper. However, the number of simulations performed there is very small, and I was not as convinced as the authors appear to be that Wolfe's suggested χ^2_{2k} approximation is an obvious failure. Consequently I have performed a number of simulations of my own, for the normal mixture case, when the null hypothesis is that the data arise from a single normal population against the alternative that they arise from a mixture of two normal components with identical covariance matrices.

Such simulations have shown that the statistic

$$L = \frac{-2}{n}.(n-k|2-2)\log l$$

has a null distribution which appears to be very well approximated by a χ^2_{2k} distribution. Some relevant results are shown in Table D1.

Table D1

				Percentage of L values	
		Mean value	*Variance*	*Exceeding* 5% *and* 1% *point of* χ^2_{2k}	
	k	*of L*	*of L*	(5%)	(1%)
$n = 50$	1	2·2	4·0	5·6	1·0
	2	4·0	8·4	4·6	1·6
	3	6·2	11·6	4·0	1·2
	4	8·2	14·5	3·8	0·8
	5	9·8	17·0	2·4	0·4

Each entry is based on the results of 500 simulations.

Passing on now to the normal mixture model discussed in Section 5.2, I would like to express some sympathy with the authors in the problems they encountered in trying to fit this model to such a large set of data. From my own experience with far smaller data sets, I suspect that the "huge amounts of time" mentioned by the authors really were huge amounts of time! Perhaps the authors could have considered fitting the model to a subset of the pupils, and then using the remainder in some form of validation process. An interesting point made by Binder in respect of such models is that the estimated posterior probabilities generally used to allocate individuals to clusters may be inappropriate because of the bias and possibly large variances of the maximum likelihood estimates of the parameters.

Finally I was encouraged to note that for each of the latent classes described in Table 1 there was a high probability that pupils were expected to know tables by heart. Perhaps this will help to convince my

10-year-old daughter that she has not been especially picked upon simply because her teacher would like her to know that 11 × 12 is *exactly* 132 rather than somewhere between 127 and 135. On reflection however, perhaps this indicates that she may have a future as a statistician!

Mr Ian Plewis (Thomas Coram Research Unit, University of London Institute of Education): Murray Aitkin and his colleagues have given us an interesting reanalysis of a controversial piece of educational research. Their discussion of variance component models will, I am sure, lead to more satisfactory analyses of designs of this kind which are common in, but are not confined to, education. However, I still have serious reservations about the techniques of clustering described in Section 3. The problems of deciding how many latent classes are needed and the difficulties posed by local maxima for solutions of more than two classes, when combined with the unlikely assumption of conditional independence, produce a degree of arbitrariness which I find hard to accept. The two- and three-class solutions, although not the local maximum solution given in Table 4, divide the sample into roughly equal groups—should I be worried about this? Would it be possible to validate the solution on the 3rd-year teachers and, incidentally, why were there so many more 3rd-year teachers in the study (790 against 468 used in this analysis)?

Dr D. B. Rubin (Datametrics Research Inc., USA): The authors are to be commended for undertaking such a massive and careful statistical analysis of an important data set. Since we learn about inferential tools primarily through seeing the consequences of their use with real data, attacking real data leads to better theoretical understanding of inferential issues, development of improved statistical tools and definition of new statistical problems needing solutions. Two generally relevant statistical issues raised in this paper are: treating parameters as random variables and summarizing inferential precision by likelihood criteria.

In Section 4.2 the authors seem to imply that the treatment of parameters as random variables should be restricted to cases in which units represented by parameters are a random sample from a larger population and the focus is on the estimation of variation among these units in the population. But I believe lessons from statistical theory and application teach us that treating parameters as random variables can be quite generally a good idea.

For one example, suppose that from the teacher data we had wished to estimate an effect for each teacher, with the entire population of teachers of interest included in the sample. One method would be to use the same normal variance-component prior distribution for these effects as used by the authors, and estimate the effects from their normal posterior distribution given maximum likelihood estimates of variance components. This procedure is computationally equivalent to the authors' method since their *EM* algorithm produces the posterior mean and variance of teacher effects given estimated variance components at each *E*-step. This empirical Bayes (cf. Efron and Morris, 1975) approach would probably lead to better answers than treating the teacher effects as fixed unknown quantities (without a prior distribution) to be estimated by maximum likelihood. A recent educational example in which such an empirical Bayes approach produced better real world answers in the sense of better predictions of future observable events is given in Rubin (1980).

In the above example, the parameters of primary interest were treated as random variables and estimated by their posterior distribution given maximum likelihood estimates of nuisance parameters. In other examples, nuisance parameters are treated as random variables, integrated out, and parameters of primary interest are estimated from the resultant likelihood. Standard cases when treating nuisance parameters as random variables produces more acceptable answers include the estimation of residual variance in the linear model and estimation of variance components in variance components models. As the authors point out in Section 4.7, "*REML*" estimation of variance components is easily handled by *EM*; Dempster, Rubin, Tsutakawa (1981) describe *EM* algorithms for such estimation treating the fixed parameters as random variables with flat prior distributions.

The estimation of scale parameters after integrating over location parameters, although desirable in general, seems computationally difficult in some commonly used models. One example is estimation by iteratively reweighted least squares regression viewed as maximum likelihood under a *t*-family specification for residuals (Dempster, Laird and Rubin, 1980); here, the residual scale would preferably be estimated from its posterior distribution (or likelihood) having integrated over regression parameters. A similar example where integration over regression parameters is desirable but apparently difficult is factor analysis. The usual maximum likelihood estimation of residual variances in this model ignores the

REML-like adjustment to these estimates that would be obtained by first integrating over the regression coefficients (factor loadings), and so *ML*-factor analyses persistently underestimate variances.

My second and final point concerns the need to avoid summarizing complicated data analyses solely by a maximum likelihood point estimate, $\hat{\theta}$, and the associated second derivative of the log likelihood, $D^2(\hat{\theta})$. As the authors point out in Section 6, non-convex likelihoods are not uncommon with complicated models, and then $\hat{\theta}$ and $D^2(\hat{\theta})$ alone are of limited inferential interest. Even if the likelihood is convex and basically symmetric about $\hat{\theta}$, I expect that $\hat{\theta}$ and $D^2(\theta)$ alone often give a poor indication of the range of plausible values of θ because $D^2(\hat{\theta})$ gives a poor indication of the global spread of the likelihood. Even approximate posterior percentiles of θ, obtained perhaps by Monte Carlo methods, would often provide the basis for more realistic inferences for θ. Consequently, I suspect that the lack of automatic calculation of the second derivative by *EM* is generally not a "major [statistical] disadvantage" (Section 6) relative to other algorithms, and may in fact lead to the examination of more satisfying measures of inferential precision. I think the authors are in general agreement with me on this point, but it is important and worth emphasizing.

Mr Stephen Simpson (Bradford Metropolitan Council): The authors' careful reanalysis makes advances on two statistical fronts. In the use of a probability model for cluster analysis, the arbitrary nature of the number of latent classes and the unlikely independence assumption for items within classes (Section 3), I feel are unfortunate dampers on its success. The proposed modelling of a continuum of teaching styles—at the same time conceptually more useful and more respectful to teachers—might overcome these problems.

On the clustered population front, however, the aptness of the variance component mixed model for hypothesis testing is not in doubt. In studies of a similar nature, ones that relate teacher characteristics to pupil performance, it seems to imply that a useful sample design would take pupils of more than one teacher from each of several schools; by thus minimizing between-teacher variance—taking from it the between-*school* variance component in the analysis—teacher differences would be more sensitively assessed.

Dr D. M. Titterington (University of Glasgow): I very much welcome the raising of the problem of testing for the number of components in a mixture, discussed in some detail in Section 3.4. In particular, it is important to emphasize the inappropriateness of the traditional recipe for the null distribution of $-2\log l$. Other simpler examples are even more weird. Consider, for instance, the case of a mixture of two *known* densities $f_1(.)$ and $f_2(.)$, for which, for all x,

$$f(x) = \lambda f_1(x)+(1-\lambda)f_2(x) \quad (0\leqslant\lambda\leqslant 1).$$

Suppose that the null hypothesis is H_0: $\lambda = 1$, and that a random sample $x_1, \cdots, x_n$ is available. It is easy to show that the mixture log-likelihood is concave in λ, with its maximum within the parameter space at $\lambda = 1$ if

$$n^{-1}\sum_{i=1}^{n} r(x_i)<1,$$

where $r(x) = f_2(x)/f_1(x)$. Since, under H_0, $E(r(x)) = 1$, this will occur with positive probability. Indeed, given mild regularity conditions, this probability becomes, asymptotically, $\frac{1}{2}$, as $n \to \infty$. Thus, far from approaching a χ^2 random variable, $-2\log l$ is, asymptotically, zero with probability $\frac{1}{2}$, under H_0.

The latent class examples considered here do not show this extreme behaviour and it is intriguing that, as far as the expected value is concerned, the asymptotic distribution of $-2\log l$ does not look far from χ^2_{2k}. A purely empirical observation can be made that the assumption $-4\log l\sim\chi^2_{4k}$ seems to fit the variances as well. This is supported by the somewhat isolated result from Wolfe (1971) for a case with $k = 22$ and a "one class versus two classes" test, for which mean 43·02 and standard deviation 6·79 are quoted for $-2\log l$.

The Authors replied later, in writing, as follows.

We would like to thank the discussants for their helpful remarks and the interest they have shown in this paper. Since many of the points were raised by more than one speaker we will deal with general issues rather than replying to individual contributors. Most of the topics have already received some discussion in the paper so we will limit ourselves to a brief reply.

Many of the discussants have serious reservations about the latent class model. We agree with Mr Hutchison's comment that the model provides a statistically sound basis for clustering, given the assumptions of the model. We believe this to be a major step forward. The fact that, as Dr Murtagh and Dr Nicholson note, different cluster algorithms give different answers raises the obvious question of which should be believed, and why. It seems to us that, when clustering samples from a population, no cluster method is *a priori* believable without a statistical model. Difficulties remain with the conditional independence postulate, as pointed out by Dr Cran and Mr Hutchison.

Mr Hutchison asks why we did not validate the latent class model on the third-year teachers. The data on these teachers is unfortunately not on file, but we did split the fourth-year teachers into two halves and fitted separate (two-group) models to each half. The parameter estimates in each half agreed very closely.

Would style be better described by a continuum, or a multidimensional latent space? The binary factor model used by Professor Bartholomew is certainly more flexible, and his results are intriguing. We will report the maximum likelihood analysis of the two-factor model for the behaviour variables, and its application to the achievement data, at a later date (work by Ian Pate in progress). Perhaps a word of warning is in order here: the binary factor model and the classical normal factor model may also have local maxima of the likelihood function, and the m.l. estimates and the second derivatives of the log-likelihood may give a very poor representation of the shape of the likelihood function for these models as well (Rubin, personal communication).

The likelihood function for the two-component normal mixture model has very peculiar contours in λ and δ ($= \mu_2 - \mu_1$), and in our experience the behaviour described by Dr Titterington is quite common in small samples when $\delta < \sigma$, the common standard deviation. This makes us doubt the value of empirical approximate distributions for $-2 \log l$. Mr Everitt's percentage point agreement appears to become worse as k increases, and the number of parameters becomes a large fraction of the sample size.

The asymptotic χ^2 distribution for the LR test *does* apply when testing the subhypothesis $\theta_{1l} = \theta_{2l} = \dots = \theta_{kl}$ for an individual item l or a group of items. Using this test it may be shown that all but 5 of the 38 items contribute to group separation in the three-group model. The five items are, in order of omission, with their associated null hypothesis χ^2_2 values, 28 (ii) (2·84), 14 (3·25), 3 (6·21), 27 (iii), (6·51), 27 (iv) (6·15). This test also establishes clearly that using the original 19 items of TS results in a substantial loss of group separation.

Dr Murtagh raises an important question: is it the clustering method itself, or the use of only 19 variables, or both, which results in the different assignments of the 36 teachers? Table D2 gives the latent class probabilities and TS style categories for these teachers using only the 19 variables of TS.

A comparison of Tables 6 and D2 shows that (i) the latent class model with 19 variables appears to identify class membership much more positively than the model with 38 variables; (ii) class membership with 19 variables changes dramatically from 38 variables for the TS mixed and informal teachers; (iii) class membership with 19 variables does not correspond at all to the original TS analysis except for

TABLE D2

Latent class probability and TS style category using 19 TS variables

TS style:	*Formal*			*Mixed*			*Informal*		
Latent Class:	1	3	2	1	3	2	1	3	2
	99	01	—	100	—	—	—	—	100
	100	—	—	100	—	—	—	—	100
	100	—	—	100	—	—	—	100	—
	98	02	—	65	35	—	—	—	100
	100	—	—	—	—	100	—	100	—
	100	—	—	01	—	99	—	—	100
	100	—	—	—	—	100	—	100	—
	100	—	—	100	—	—	—	—	100
	98	02	—	89	—	11	—	—	100
	100	—	—	92	08	—	—	100	—
	99	01	—	—	—	100	—	100	—
	98	02	—				—	100	—
							—	100	—

formal teachers. Note that the 19 omitted variables include the important variables 2, 4, 6, 7, 11, 17, 24, 25, 26 and 27 (iv).

A point raised by several discussants is the complexity of the analysis. Could simpler answers have been obtained? Dr Cran's suggestion would substantially reduce the number of variables in the latent class model, but requires an *a priori* weighting of the items into sub-scales. In response to Professor Greenfield's question, the optimal weights for the two-class model are those given in Table 1 under the δ heading. Analogous optimal weights can be obtained for discriminating between each *pair* of classes in the three-class model.

Professor Bartholomew's comment about the inability to discriminate between the two- and three-category latent class models puzzles us. The observed marginal distribution of total formality score might also occur under a two-factor model, in which case this observed distribution would not discriminate between the two-factor and the three-latent-class model. However, it is clearly possible to discriminate between the two- and the three-latent-class models from this marginal distribution.

Miss Reid's suggestion requires an *a priori* specification of class membership. The latent class model can be viewed as a logistic discriminant model with an unobservable "response"—the class membership. The *EM* algorithm alternates logistic discrimination using the current probabilities of class membership and the calculation of conditional probabilities of class membership given the current parameter estimates.

Professor Cox's suggestion is appealing, but vague. There are difficulties in using the class as the unit of analysis, as described in Section 4.

If style is largely irrelevant to achievement, are there other variables which *are* relevant? Sex, age, type of training and years of experience are not available on file, but Bennett (personal communication) compared the three TS styles on these variables. No substantial differences among styles were found on any of these variables. There may be differences among classes, but this question remains to be investigated.

Mr Hutchison finds it hard to understand why anyone should have expected that teaching style should have made much difference to achievement, but this is surely, as he concedes, the wisdom of hindsight. The TS study was carried out precisely because the issue was so hotly debated at the time.

We would like finally to acknowledge Dr Rubin's contribution to this paper. The reanalysis has been strongly influenced by his fundamental *EM* paper with Dempster and Laird, and we have benefited personally as well from many discussions during his visits to Lancaster.

References in the Discussion

Binder, D. A. (1977). Cluster analysis under parametric models. Ph.D. Thesis, University of London.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. *Multiv. Anal.*, V, 35–37.

Dempster, A. P., Rubin, D. B. and Tsutakawa, R. K. (1981). Estimation of covariance components models. *J. Amer. Statist. Ass.*, **76**, 341–353.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalization. *J. Amer. Statist. Ass.*, **70**, 311–319.

Gray, J. (1979). Reading progress in English infant schools: some problems emerging from a study of teacher effectiveness. *Brit. Educ. Res. J.*, **5** (2), 141–157.

Keddie, N. (1971). Classroom knowledge. In *Knowledge and Control* by M. F. D. Young. London: Routledge, Kegan Paul.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45**, 325–342.

Pawson, R. (1978). Empirical explanatory studies: the case of causal modelling. *Sociol. Rev.*, **26**, 613–645.

Rubin, D. B. (1980). Using empirical Bayes techniques in the law school validity studies (with Discussion). *J. Amer. Statist. Ass.*, **75**, 801–827.

Rutter, M., Maughan, B., Mortimore, P. and Ouston, J. (1979). *Fifteen Thousand Hours*. London: Open Books.

Steedman, J. (1980). *Progress in Secondary Schools*. London: National Children's Bureau.

Wilby, P. (1981). Higher score for the new teaching. *The Sunday Times*, April 26th.

Wolfe, J. H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinomial distributions. *Tech. Bull. STB 72-2, Nav. Res. Trg. Lab.*, San Diego.