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criticisms, and the statistical issues remained unresolved, despite further discussion of the statistical aspects in an unpublished report by Satterly and Gray (1976).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

	Item	*Two-class model* Class 1	Class 2	δ	*Three-class model* 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.
δ 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

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

with

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

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

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

Then

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

where c is a proportionality constant, and hence

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

where ψ_{jl} is the *log-odds*:

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

We may write shortly

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

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_{jl}^* and λ_j 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

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

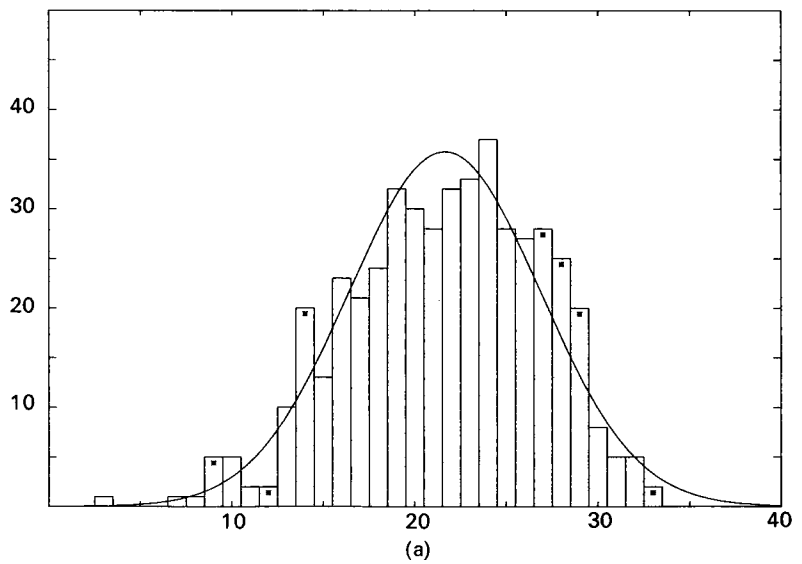
Fig. 1(a) shows the distribution of total formality score T for the 468 teachers, with the superimposed normal distribution with estimated mean $\hat{\mu} = 21{\cdot}64$ and variance $\hat{\sigma}^2 = 5{\cdot}22^2$. The overall goodness-of-fit χ^2 is 36·06, on 24 d.f., if the three smallest observations 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)

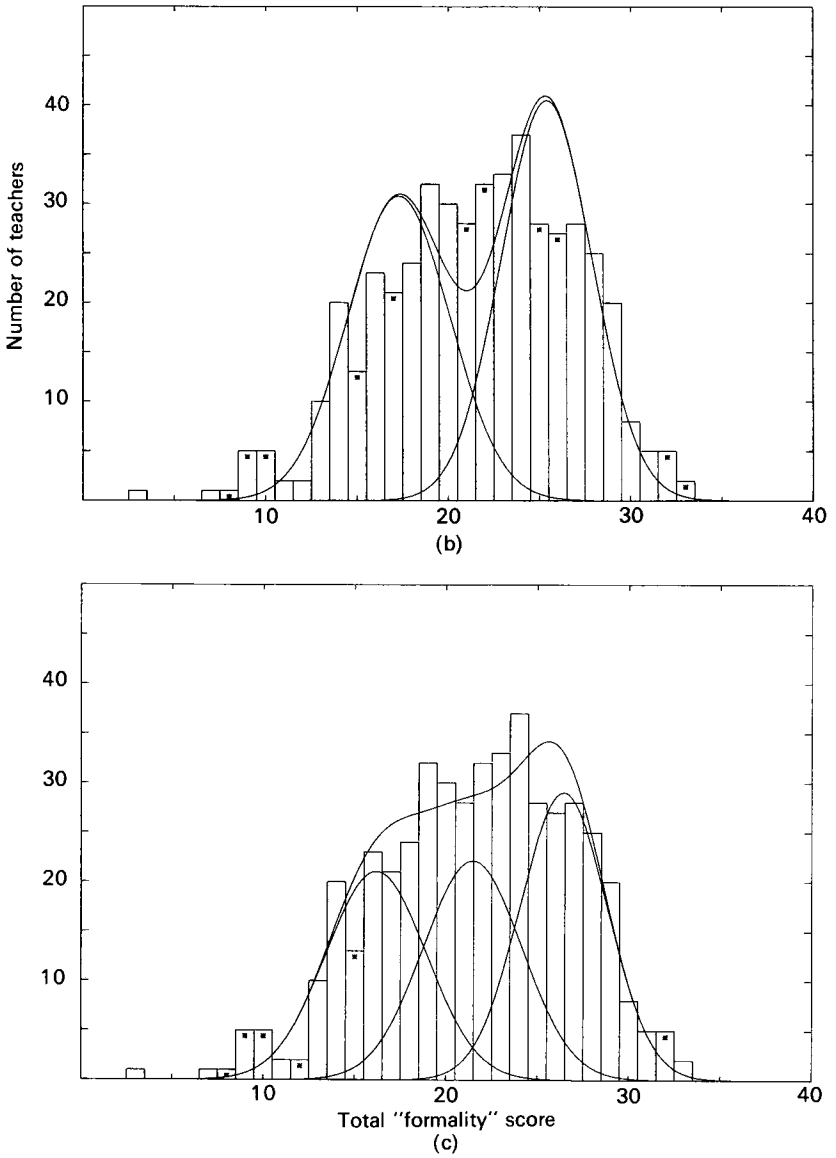


FIG. 1. (a) Homogeneous population. (b) Two classes. (c) Three classes.

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 integrated 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"	—	—	5	9	4	2	9	20	19	24	31	36	16	175
(Class 1)	(0)	(0)	(21)	(27)	(15)	(5)	(30)	(67)	(53)	(77)	(79)	(100)	(21)	
"Mixed"	1	13	11	2	7	26	19	6	14	6	8	—	31	144
(Class 3)	(3)	(41)	(46)	(6)	(27)	(69)	(63)	(20)	(39)	(20)	(21)	(0)	(39)	
"Informal"	34	19	8	22	15	10	2	4	3	1	—	—	31	149
(Class 2)	(97)	(59)	(33)	(67)	(58)	(26)	(7)	(13)	(8)	(3)	(0)	(0)	(40)	
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

Latent class:	*TS style* *Formal* 1	3	2	*Mixed* 1	3	2	*Informal* 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:

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

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

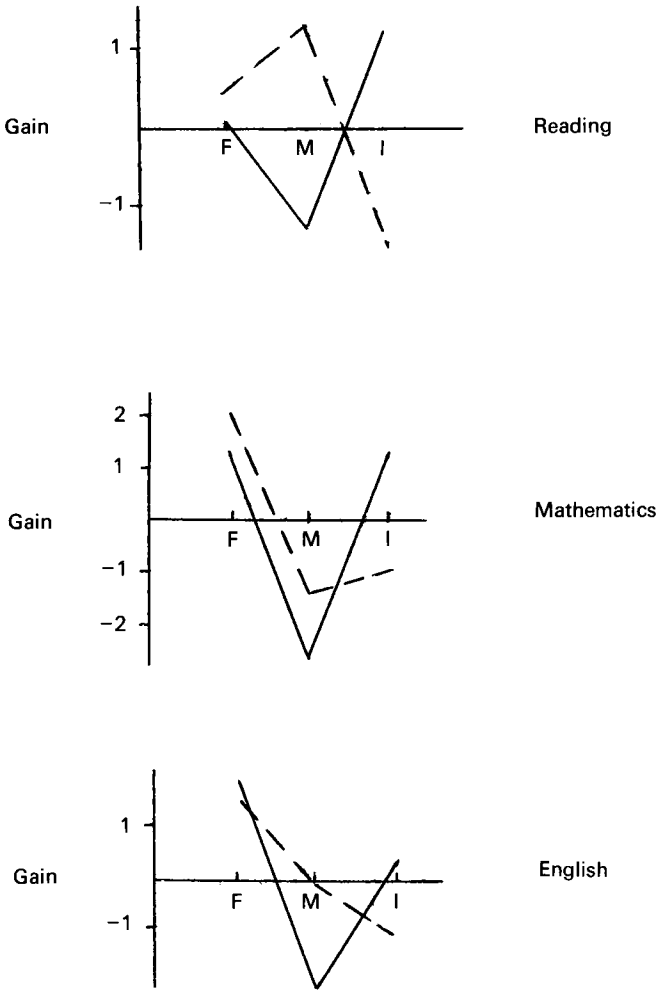


FIG. 2. Comparison of teaching style differences, TS and reanalysis. F, Formal; M, mixed; I, informal. - - -, TS; ——, reanalysis.

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 $E(\mathbf{Y}) = X\boldsymbol{\beta}$, $\mathcal{V}(\mathbf{Y}) = \sigma^2 H$, where

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

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

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

Thus the sufficient statistics involve the unknown **T** through **T**, **T**'**T** and **T**'$W'W$**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

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

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

$$P(\mathbf{X}_i = \mathbf{x}_i \mid z_{ji}) = \sum_{l=1}^{38} P(X_{il} = x_{il} \mid j, \theta_{jl}), \quad i = 1,\ldots,468,$$

$$Y_{ji'r} \mid z_{ji'}, T_{i'} \sim N(\mu + \gamma u_{ji'r} + \alpha_j z_{ji'} + T_{i'}, \sigma_E^2), \quad i' = 1,\ldots,36,$$

$$T_{i'} \sim N(O, \sigma_T^2), \quad j = 1,2,3,$$

$$P(z_{ji'} = 0) = 1 - \lambda_j, \quad r = 1,\ldots,n_i,$$

$$P(z_{ji'} = 1) = \lambda_j$$

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 $Y_{ji'r}$ 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 $\mathbf{X}$, 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 $\boldsymbol{\Sigma}$. 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, $\boldsymbol{\Sigma}$ 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 $\mathbf{X}$ on the factors $\mathbf{U}$, 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 $\mathbf{X}$ 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

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

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

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

where

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

Thus

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

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

Variable	*Range in TS data*	*Mean in TS data*	*Regression coefficients*		
			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.