Curriculum, Evaluation and Management Centre

PUBLICATION NO.6

MULTI-LEVEL MODELLING IN AN INDICATOR SYSTEM

C.T. FITZ-GIBBON

1991

Schools, Classrooms, and Pupils

International Studies of Schooling from a Multilevel Perspective

Edited by

Stephen W. Raudenbush

College of Education
Michigan State University
East Lansing, Michigan

J. Douglas Willms

Centre for Policy Studies in Education
University of British Columbia
Vancouver, British Columbia, Canada
and Centre for Educational Sociology
University of Edinburgh
Edinburgh, Scotland

1991

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
San Diego New York Boston London Sydney Tokyo Toronto

Chapter 6

Multilevel Modelling in an Indicator System

Carol T. Fitz-Gibbon[1]
University of Newcastle Upon Tyne

Purposes and practical considerations enter into choices of statistical analysis methods. This paper considers the application of multilevel modelling to data from a Performance Monitoring system in the United Kingdom known as the A-level Information System (ALIS). The practical question concerned the extent to which it was important to use this more sophisticated model rather than simply using standard OLS regressions, given the kind of data dealt with in this particular monitoring system. We begin with a consideration of dependent variables (outcome indicators) and present an example of their apparent relationship to a process variable. The example is a springboard for a brief discussion of the purpose of a monitoring system. The results of applying ML2[2] to the analysis of four dependent variables are then presented. The paper concludes with a discussion of the limitations of the data and an attempt to assess the role of multilevel modelling for this particular indicator system. Finally, a suggestion is made that a desirable piece of statistical software might be one delivering rapid-fire graphical bootstrapping.

One of the aims of the A-level monitoring system is to report to schools how their students' examination results compared with those of similar students in other institutions, that is, to provide what might be called "fair performance indicators" or "school effectiveness indices" (SEIs). The achievement measure used in this system is quite different from that used in many other studies.

Adequate and Inadequate Measures of Achievement

In some educational systems there is a grave shortage of good dependent variables for school achievement. Willms (1985) highlighted the fact that in the High School and Beyond data, collected in U.S. high schools, "academic growth in science" was based on a 20-item test on which the average growth was less than one item over two years of instruction. Other "curriculum specific tests" had even fewer items, and growth in terms

[1]Support for the A-Level Information System, received from the Department of Education and Science and seven LEAs, is gratefully acknowledged, as is the contribution made by discussions with LEA and school personnel during the years of development. Comments on this paper by the editors were exceedingly helpful.

[2] The multilevel analysis employed ML2 1.0 (Rasbash, Prosser, and Goldstein, 1988).

of effect sizes (using standard deviations of the posttest) ranged from .06 to .20 over two years. Such tests can hardly do justice to two years of learning and instruction. Referring to the "virtually zero growth" of students who had been in the top half on the pretest, Willms summarized:

> . . . We have been estimating sector differences with tests that show virtually zero growth for half of the population and a very small amount of growth for the other half of the population, much of which is achieved by students who did not attend school during the intervention period.

The problems of poor measures of school achievement (inadequate content and construct validity) have been recognized in the US for many years, as illustrated by Carver's (1975) critique of achievement tests used in the Coleman reports. See also Dyer (1968) and Jencks (1972). Tests labeled as "achievement" tests do not necessarily measure the kind of achievement for which a school can be held responsible.

Murnane (1987), reviewing the history of education indicators in the US, arrived at this conclusion:

> Of particular value would be selective use of tests that examine how students solve problems, and that do not restrict answers to a predetermined set of possibilities.

He argued that the information to be gained from such tests would justify "the significant cost of developing, administering and correcting open-ended response tests on a selective basis."

Perhaps the UK already has the kind of tests which Murnane suggests, in its externally set and graded examinations called O-levels (ordinary level examinations taken at age 16) and A-levels (advanced level examinations taken by about 17 to 20 percent of 18-year-olds). For these external examinations teachers teach to a published syllabus, and students are assessed by complex and multi-faceted examinations based on this syllabus. The examinations for each academic subject are professionally developed and graded by independent organizations known as examination boards.

Subsequent to the publication of examination results, the Boards even publish advice on the kinds of errors found in the students' work so that teachers can improve the delivery of instruction in subsequent years. Thus, there is a *tight link between instruction and assessment*, and the effects of instructional strategies might, in such circumstances, be detectable.

If such a test were sensitive to instructional effects, then the monitoring of outcomes on this test might inform practice. Instructionally relevant hypotheses can be generated from the data in the indicator system known as ALIS (the A-level Information System) by relating outcome indicators to instructional processes.

An Example of Two Indicators Related to a Process Variable

If the A-level examination is sensitive to instructional effects, then we should be able to locate some relationships between teaching processes and outcome variables. Two examples of such relationships are presented here based on data from 461 students taking the Chemistry exam in the summer of 1988.

Table 1

a) Attitude to Subject Related to Frequency of Essays, Chemistry, 1988

Source of Variation	Sum of Squares	DF	Mean Square	F	p-value
Essays	5.62	4	1.406	1.91	.10
Sex	5.61	1	5.608	7.61	.006
Essays by Sex	8.57	4	2.143	3.13	.001
Residual	331.10	451	.736		
Total	352.87	460	.767		

b) A-grade Examination residuals related to frequency of Essays in Chemistry, 1988

Essays	18.7	4	4.67	2.86	.04
Sex	6.5	1	8.48	2.47	.06
Essays by Sex	5.0	4	1.25	.66	.62
Residual	854.7	451	1.89		
Total	886.8	460	1.93		

In Figure 1a the dependent variable (on the vertical axis) is *attitude-to-the-subject*, in this case Chemistry. Mean scores on this measure have been graphed against students' reports of how often they wrote essays in their A-level Chemistry course, with male and female responses graphed separately. It can be seen that male students who were most positive about Chemistry were those who reported writing essays once a term. Any demand for more frequent essays seemed to have been associated with less positive attitudes. Female students, on the other hand, showed less positive attitudes except for the group who were set essays fortnightly or more often.

Of course, these were correlational data and could not be interpreted in terms of causal relationships. Nevertheless they suggested an interesting hypothesis about the possibility of setting essays to improve girls' attitudes to Chemistry, a hypothesis which is consistent with research suggesting that girls like to see science in context, with its history and implications as well as its conclusions and details (Kelly and Smail, 1984). Essays would allow this holistic approach. Of course, gender might not have been the "real" explanatory variable. Perhaps the graph would have shown even stronger effects if some cognitive-set variables had been used instead of gender: for example "need for relevance" as opposed to "interest in abstract activities."

However, students' attitudes to the subject might be regarded as not nearly as important as getting good examination results. What, then, of the examination results for the same students? Figure 1b relates the same process variable, as reported by students, to examination data in the form of residuals computed using ordinary least squares regression to 'control for' prior achievement. (The ANOVA tables for both graphs are provided in Table 1). Happily, the implication of this graph was also that setting essays once a fortnight might be a good teaching technique: residuals were a maximum for both males and females reporting this frequency of essays.

Of course, the "essays" may in fact be merely an indicator of some other variable, such as a hard-working teacher who is willing to grade essays every

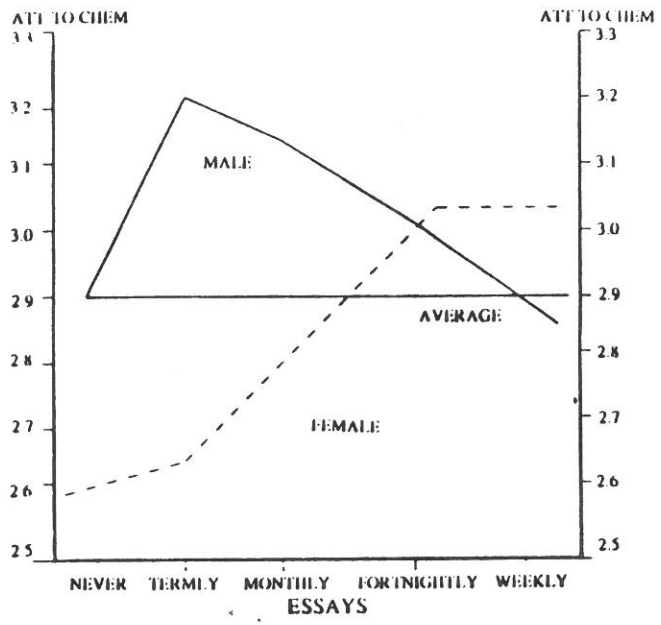

Figure 1a. Attitude to chemistry related to the reported frequency of essays being set in chemistry classes.

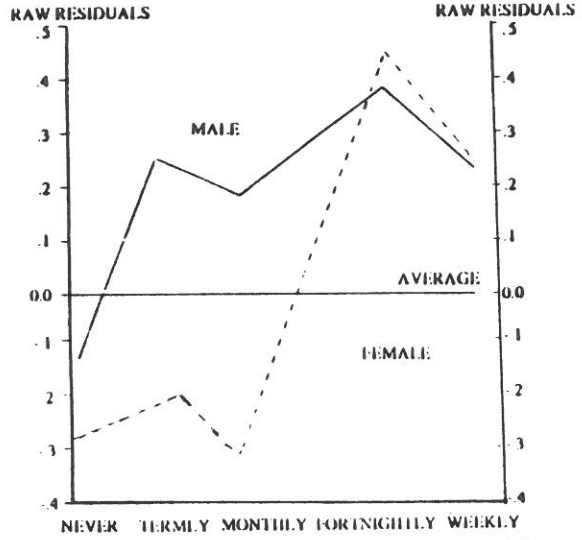

Figure 1b. Examination residuals in chemistry related to the reported frequency of essays being set in chemistry classes.

other weekend. On the other hand, setting essays fits in with research on "generative learning" (Wittrock, Marks and Doctorow, 1978), which might be roughly summarized as showing that students learn best from work which they have to organise for themselves. Whatever the explanation, these graphs suggested some testable hypotheses and were, incidentally, quite unexpected. *Essays* as a process variable had been in the questionnaire largely with the teaching of arts and humanities in mind, not chemistry.

The relationships shown in Figure 1 might not show up next year; they might have been dependent upon a small subset of atypical cases, and might not be replicable no matter how statistically significant they were on this particular occasion. More on this at the end of the paper.

The point of this example is to illustrate the idea that a monitoring system might provide practitioners with feedback which is relevant to the decisions which they make about how to deliver education. If a monitoring system can not only indicate the level of results each year in each school, but can also relate these findings to processes, the hope is that education might gradually be improved. Such a system would provide feedback about alterable variables (Bloom, 1979).

The A-Level Information System

From its inception in 1983, the A-level monitoring system has collected student-level data. The students may be viewed as nested within departments rather than schools because examination results for each academic subject are analyzed separately. In view of pressures towards inter-school competitiveness, it is perhaps quite important that ALIS does not produce *school* indicators, but only *departmental* indicators.

Some Local Education Authorities ("LEA's," similar to U.S. school districts) use the information system purely for formative, improvement-oriented, purposes. These LEAs send the data directly to each school using code-names to preserve confidentiality between schools. Each school sees all the data but knows only its own code-name. In contrast, other LEAs want to know the code-names in order to compare the performances of the schools for which they are resposible.

The Feedback from the Monitoring

Each school receives 33 reports per year: three reports for each of eleven subjects. The first report concerns the examination performance of students taking the exam in each subject. The second report concerns attitudes, and the third report links process variables to examination residuals and attitude scores. The tables in these reports typically show school means in rank order.

In reading the reports from ALIS, heads of schools and departments are urged to compare like with like by locating schools like their own on the basis of the tables showing various intake characteristics and then to see how their school or department fared on output measures in comparison with these similar schools or departments. Attention is not to be focused on the rank order.

Multilevel Modelling in an Indicator System

Prior to the advent of computer programs for multilevel modelling, the ALIS project relied on student-level Ordinary Least Squares (OLS) regression equations. The best predictor of A-level examination results was a measure of prior achievement: *O-grade*, the student's Grade Point Average in the O-level examinations. For each school, the mean of the student-level residuals controlling for *O-grade* was taken as one of several "Fair Performance Indicators". That is, it represented a "School Effectiveness Index" (SEI) for that academic subject.

Variables and Sample

For this pilot exploration, a small subset of variables out of the more than 250 derived from the questionnaires, ability tests, and examination data were used. They are listed below:

Outcomes

A-Grade — A-level examination score for each of four subjects (chemistry, geography, French, and mathematics)

Attitude to Subject

Attitude to School

Aspiration — level of aspiration for higher education

Predictors

O-Grade — prior achievement: average O-level grade

Ability — ability score on International Test of Developed Ability

Sex — gender (Male was scored 1 and female 2)

SES — Socioeconomic status based on Registrar General's Scale of occupation of head of household (6=professional; 5=semi-professional; 4=white collar; 3=skilled manual; 2=partly skilled; 1=unskilled)

The samples for the analysis were derived by selecting a set of variables to be investigated (displayed above), and using listwise deletion of missing cases. Table 2 displays some basic statistics for the A-level subjects. Note that *O-grade* correlated consistently and substantially with the *A-grade* whereas *ability* was not an effective predictor in language-based subjects. The ratio of males to females varied as expected between languages and sciences, with the smallest proportion of females in mathematics.

Exploring Possible Models

Because there were more data for mathematics than for other subjects, the mathematics file was used to explore possible models. Table 3 presents results for *A-grade* (math) as the dependent variable. Results for the null model, with no explanatory variables, indicated that there was significant variation among schools, amounting to about 11 percent of the total variance: (0.45/4.14)•100. Both the among-school and the pupil-level variances were partly explained by the variables in the specified model (*O-grade, ability, sex* and *mean O-grade* for the school) so that the total variance was reduced to 2.35. Thus R^2 for the total variation was .43. For school variation, R^2 was (3.69-2.04)/3.69 = 0.45.

Table 2

Descriptive Statistics for the Academic Subjects

	Mean	SD	Correlations			
			A-Grade	O-Grade	Ability	Sex
Chemistry (n=394, j=26)						
A-Grade	2.50	1.91				
O-Grade	5.83	.61	.63			
Ability	71.60	12.68	.47	.38		
Sex	1.43	.49	-.10	.09	-.24	
SES	4.27	1.31	.13	.22	.06	.03
Geography (n=310, j=28)						
A-Grade	1.71	1.76				
O-Grade	5.37	.53	.45			
Ability	61.59	10.92	.23	.22		
Sex	1.43	.50	.10	.14	-.08	
SES	4.12	1.27	.04	.10	.01	.02
French (n=152, j=27)						
A-Grade	2.15	1.80				
O-Grade	5.83	.62	.66			
Ability	60.34	11.60	.24	.29		
Sex	1.81	.39	-.13	-.03	-.18	
SES	4.15	1.24	.18	.20	-.04	.02
Mathematics (n=591, j=30)						
A-Grade	2.07	2.03				
O-Grade	5.74	.58	.55			
Ability	73.29	11.51	.51	.35		
Sex	1.31	.46	.03	.21	-.13	
SES	4.29	1.26	.20	.20	.08	-.02

Slope Heterogeneity

If the within-school regression slopes were not parallel, the model would need to include terms to model slopes, and school effects would need to be specified separately for pupils who differed on the predictors. To detect possible slope heterogeneity the slopes were allowed to vary randomly. However, no significant slope variation was found. The model used was therefore a "variance components model." The lack of statistically significant heterogeneity in the slopes was expected because of lack of statistical power resulting from the small number of level-two units (departments) for which there were complete data and the very small samples in some of those units.

Context Effects

A "context" effect for a variable which is measured at level-one can be defined as the effect of some level-two aggregate statistic for the variable, over and above its effect as a level-1 variable. The context effect is a measure of the effect of the composition of the groups. Thus, in the following equation X_{ij} is included as a level-one predictor and the mean of X_{ij}, denoted $\bar{X}_j$, is the level-two aggregated statistic. The context effect is γ.

$$Y_{ij} = \beta_{0j} + \beta_w X_{ij} + \gamma \bar{X}_j + u_j + \epsilon_{ij}$$

Table 3

Exploring the Largest Available Subsample: Outcome is Maths O-Grade

	Null Model		Specified Model	
	Effect	SE	Effect	SE
Effects of Student-Level Variables:				
Intercept	1.96	(.14)	-11.44**	(3.27)
O-Grade			1.45**	(.10)
Ability			.06**	(.005)
Sex			-.14	(.11)
Effects of School-Level Variables:				
Mean O-Grade			2.34**	(.57)
Model Statistics				
Intra-School Correlation	.11		.13	
Variance within schools	3.09		2.04	
Variance among schools	.45		.31	
Total unexplained variance	4.14		2.35	

Note: ** t > 3.00

The mean of X_{ij} was used, although the median or standard deviation or some other aggregate statistic could be used. When the level-one variables are centered about their school means, the equation becomes[3]:

$$Y_{ij} = \mu_j + \beta_w(X_{ij}-\bar{X}_j) + \beta_b\bar{X}_j + u_j + \epsilon_{ij}$$

The context effect is therefore represented in the second equation by the difference β_b-β_w (Raudenbush, 1989a). The test for a contrast effect given by Raudenbush (1989b) showed that the context effect was not statistically significant, as indeed had been shown in the results for the uncentered model. The choices of centering on group means or the grand mean did not affect the amounts of variance accounted for.

A Model For All Seasons

Two considerations were important in selecting a single model to be applied across the board, that is, for all four subjects and each of the four dependent variables. First, there was the practical concern that in a working indicator system, in which schools appreciate rapid feedback, it would certainly be more efficient to apply the same model for each subject and each dependent variable than to experiment with different models for 44 situations (11 possible subjects x 4 dependent variables). Secondly, it was of

[3] In the present analysis, the student-level predictors *O-Grade* and *Ability* were centered about their school means, and the predictor means were used as predictors at level two. In situations where there is no contextual effect, and no heterogeneity of regression, the use of X variables centered on their grand mean rather than group means would probably have been preferable (Raudenbush, 1989b, p. 12). However, the differences between centering on group means or on the grand mean should be insubstantial for the data under consideration when the group means are also included as predictors in the model.

Table 4

A-Level Grades related to Five Predictors

	Estimate	SE	Z
Chemistry			
Constant	-9.453		
Student-Level Coefficients			
O-Level GPA	1.825	.126	14.48
Ability	.031	.006	5.16
Sex	-.398	.143	-2.78
School-Level Coefficients			
Mean O-Level GPA	1.456	.585	2.50
Mean SES	.199	.278	.72
Intra-School Correlation	.15		
Variance Within Schools	1.69		
Variance Among Schools	.31		
Geography			
Constant	-9.184		
Student-Level Coefficients			
O-Level GPA	1.381	.168	8.22
Ability	.023	.008	2.88
Sex	.143	.178	.80
School-Level Coefficients			
Mean O-Level GPA	1.800	.698	2.58
Mean SES	.099	.289	-.34
Intra-School Correlation	.15		
Variance Within Schools	2.05		
Variance Among Schools	.36		
French			
Constant	-12.53		
Student-Level Coefficients			
O-Level GPA	1.709	.191	8.95
Ability	.014	.009	1.56
Sex	-.527	.278	-1.90
School-Level Coefficients			
Mean O-Level GPA	1.918	.426	4.50
Mean SES	.312	.258	1.21
Intra-School Correlation	.16		
Variance Within Schools	1.41		
Variance Among Schools	.27		
Mathematics			
Constant	-10.200		
Student-Level Coefficients			
O-Level GPA	1.439	.119	12.09
Ability	.063	.006	10.50
Sex	-.082	.141	-.58
School-Level Coefficients			
Mean O-Level GPA	1.863	.590	3.16
Mean SES	.338	.341	.99
Intra-School Correlation	.13		
Variance Within Schools	2.11		
Variance Among Schools	.31		

Note: All student-level variables were centred.

interest to compare the contributions of each of the predictor variables included in the model.[4]

Table 4 presents the results of applying the model to the *A-grade* in four subjects: chemistry, geography, French, and mathematics. *SES* did not have a statistically significant effect in any of the four equations.

We see from these data that a rather large proportion of variance was attributable to schools. The average value of about 15 percent was larger than the 10 percent often reported in the school effects literature (for example, most recently, Smith and Tomlinson, 1989).

An immediate question concerns a possible source of inflation of the differences between schools: the methods of collecting data in the schools. If there were differences in the conditions in which data (ability measures and questionnaire responses) were collected, then any effects arising from these administration conditions would be completely confounded with school effects. In the ALIS project, strenuous efforts are made to standardize the conditions by sending trained data-collectors from the university to collect the data using tape-recorded instructions and timing (Fitz-Gibbon, 1985). These precautions increase the likelihood that the larger effects attributable to schools might be genuine. If they are genuine, then this finding reflects the tight link mentioned earlier between instruction and the dependent variable: *A-grades* may well be particularly sensitive to instructional effects.

The model applied to examination outcomes was applied to the three other dependent variables: *attitude to school, attitude to subject* and *aspirations* for higher education. A subsequent analysis might profitably use a three-level model. The four dependent variables within each subject might have been treated as multiple outcomes (level one) nested within students (level two) nested within school departments (level three). This would allow the intercorrelations among the dependent variables to be taken into account and would allow more precise estimation of the indicators.

The Differences Made By The Use Of Multilevel Approaches

The ALIS indicator system has been providing analyses based on multiple regression using SPSS*. In the absence of slope heterogeneity there is certainly a temptation to continue to do so rather than to change to multilevel modelling, adding an additional layer of software to the task of preparing reports to meet frequent deadlines. It is therefore important to consider what substantive differences might result from using SPSS* and the analysis presented above using ML2. Only tentative answers can be provided on the basis of the 1988 data.

[4] Obtaining 808 cases for mathematics could be accomplished only by not including any measure of socio-economic status (*SES*). In the questionnaire data, the *SES* was often missing or could not be coded. Thus, when *SES* was included in the mathematics data and listwise deletion used, the sample dropped to 591 complete cases. It was felt necessary to include *SES* because of the possibility that, although it had no statistically significant effect on mathematics *A-Grade*, it might nevertheless affect other dependent variables such as attitudinal ones, or have effects on achievement in subjects other than mathematics. *Mean SES* was included as a level-2 variable in order to explore a model which could ignore some of the missing level-1 measure of *SES* and as a proxy for the kind of school attended.

Table 5

Mean Residuals (SEIs) from ML2 and SPSS: Mathematics*

School ID	Size	ML2 Residual	SPSS* Residual	Difference
30	16	-1.538	-2.060	-0.52
48	22	-.629	-.771	-0.14
43	11	-.447	-.332	0.11
46	8	-.425	-.537	-0.11
19	18	-.402	-.443	-0.04
42	40	-.400	-.533	-0.13
8	40	-.354	-.559	-0.20
49	10	-.347	-.340	0.00
47	19	-.345	-.825	-0.48
34	2	-.312	-1.696	-1.38
31	14	-.278	.035	0.31
26	12	-.264	-.507	-0.24
23	5	-.106	-.283	-0.17
17	17	-.100	-.083	0.01
40	9	-.072	-.140	-0.06
24	7	-.068	-.352	-0.28
28	5	-.030	-.753	-0.72
50	51	-.002	.267	0.26
32	6	.019	.791	0.77
16	17	.040	.221	0.18
21	7	.067	-.402	-0.46
9	45	.085	.018	-0.06
20	6	.226	.167	-0.05
22	7	.252	.472	0.22
15	28	.304	.687	0.38
39	7	.307.	.638	0.33
6	41	.357	.419	0.06
4	18	.627	.549	-0.07
18	51	.660	.532	-0.12
7	52	.861	.878	0.01

To compare methods, OLS regression was used to compute residuals. The schools were then used as levels of the independent variable in a one-way analysis of variance on the residuals. The results were quite similar to those produced by ML2. This is more or less to be expected since there was no modelling of slopes and level 2 variables were generally not influential.

A critical outcome of the analyses, however, is the nature of the school residuals. The analysis provided by ML2 will 'shrink' residuals derived from small samples, on Bayesian grounds that the overall pattern should be allowed to have more influence the smaller the within-school sample. The shrinkage factor, according to Goldstein (1987 p.21) is

$$n_i \sigma_u^2 (n_i \sigma_u^2 + \sigma_e^2)^{-1}$$

The shrinkage will be related therefore to the sample size and to the proportion of variance the model attributes to level-two units. The shrinkage for the mathematics sample is shown in a listing of the residuals and the within-school sample sizes in Table 5.

Table 6

Correlations between SEIs from ML2 and OLS

Subject	A-Grade	Attitude to School	Attitude to Subject	Aspiration
Chemistry	.94	.77	.78	.90
Geography	.97	.91	.94	.44
Mathematics	.84	.77	.79	.84

In Table 6 the correlations between the OLS and ML2 residuals are reported for each dependent variable and subject. (French is omitted because of a lack of school variance on the non-cognitive outcomes.) These correlations had a median value of .90 and all but one were above .77. What differences there were in the residuals resulted not only from the different analysis procedures but also from the different models used in this essentially pragmatic comparison. All variables (including SES) were used at the pupil level in the OLS model. The ML2 model used two school-level variables and these were not used in the SPSSX analysis. Also, greater OLS-ML2 differences were associated with smaller samples (see the equation above). When group sizes exceeded 30, little shrinkage occurred so that the OLS-ML2 differences were generally less than .3.

Small samples are inevitable in this particular data because it is usually only colleges of further education and tertiary colleges which have large numbers enrolled for each A-level subject, and, with declining numbers of 18-year-olds, the situation may get worse. The shrinkage factor can in fact be interpreted as the ratio of true variance to total variance and thus the "reliability" of the school effectiveness index (prior to shrinkage) (Raudenbush, 1989b). Using typical values from the tables we arrive at the following reliabilities:

Sample Size	Reliability
5	.49
10	.65
30	.85
50	.90

These values suggest that caution which must be urged in interpreting school effectiveness indices based on small samples. The index might be the best estimate for that data, but it must be interpreted tentatively. Only steady trends year after year could provide a basis for strong inferences.

In practice the exact sizes of the residuals have very little meaning in the indicator system. Schools and colleges are mainly looking at their residuals compared with institutions of similar intake. They are also watching their residuals from year to year. However, given the differential shrinkage due to sample size it would seem to be desirable, if multilevel modelling were used, to provide comparisons only between institutions of similar sizes. At present the sample size for each institution is not reported because this information would allow institutions to be identified, and strong precautions are taken in the information system to guard anonymity.

In summary, use of multilevel modelling, then, gave fairly similar rank orders but with some large differences in magnitudes of school level residuals (which are school effectiveness indices) for small samples.

Should shrinkage be used in a working indicator system to obtain the shrinkage adjustment? The following experience bears on this question. In some feedback to schools, samples containing fewer than five cases were removed because it seemed unfair to report the data of such a small sample. School personnel objected, though, to this failure to provide the indicator for group sizes of less than five. Schools wanted to see the result in the context of the overall data and felt well able to take account, for themselves, of the fact that there might have been only a few cases. Moreover, some teachers in each school followed the OLS analysis with ease. Using Bayesian methods may make them feel the data were further from their understanding and possibly less acceptable. Moreover, shrinkage might actually obscure true changes in school effectiveness. Suppose that a poor result occurred one year but that substantial improvement occurred the next year. The multi-level indicator might imply a poorer result if less shrinkage were applied to the second measure due to a change in sample size and a value closer to the norm. This could be annoying. Furthermore:

> Although the shrinkage estimates are generally *more accurate* than estimates without shrinkage (their average distance from the true score is smaller than is that of the non-shrinkage estimates), they are also *biased*. Suppose a school serving students with low prior achievement were especially effective. In this case, it would have its score "pulled" toward the expected value of schools with children having low prior achievement. That is, it would have its effectiveness score "pulled" *downward*, in the "socially' expected direction, demonstrating a kind of statistical self-fulfilling prophecy! (On the other hand a similar school doing very badly would be pulled up.) (Raudenbush, 1989a)

When using ML modelling a small sample size provides protection against being assigned a strongly positive or negative indicator and yet it is in small samples that schools are most aware of how good or poor the results are. They are able to make their own allowances for small samples but would, I believe, prefer to see unshrunk data.

Shrinkage might be desirable for a specific research agenda or in a longitudinal analysis, and should make indicators appear more stable over time. However, face validity may be more important in feeding results back to schools each year.

As already mentioned, one solution to these problems might be to enable schools to restrict the comparisons they make even further: not only to schools with similar intakes but also to schools with similar sizes, so that the shrinkage factors would be similar. However, a problem here is that only *categories* of size could be provided. It is a feature of the ALIS reports, which cannot be changed, that group sizes are not reported. To report group sizes would provide a way in which individual schools could be identified.

Table 7

The Consistency of School Effects on the Same Pupils: Correlations for ML2 SEIs below the diagonal and OLS SEIs above the diagonal

	A-Level	Attitude to School	Attitude to Subject	Aspiration
Chemistry				
A-Level	----	.19	.38	.36
Attitude to School	.31	----	-.04	.33
Attitude to Subject	.53	.02	----	-.23
Aspirations	.50	.37	.12	----
Geography				
A-Level	----	-.14	.21	-.04
Attitude to School	-.09	----	.21	.04
Attitude to Subject	.41	.30	----	.01
Aspirations	-.05	.27	.12	----
Mathematics				
A-Level	----	-.17	.44	.44
Attitude to School	.26	----	.09	.06
Attitude to Subject	.39	.48	----	.24
Aspirations	.48	.13	.01	----

Given this kind of conflict between models, a possible solution is to provide both analyses for the all-important achievement data. Residuals for these data are already reported in four different combinations (reflecting the use of different control variables). One more table might be quite acceptable. For purposes other than reporting back to the school, the multilevel residuals might be important. For example, it might be as well to use the multilevel tables for reports at the LEA level where they are simply keeping a watching brief rather than looking in detail at each result. And for research purposes, especially in looking at the changes across time, the multilevel models may be important because of their capacity to distinguish between "true" parameter variance and sampling variance. Correlations with process variables might also be improved by the use of ML residuals.

As is stressed in the ALIS reports, researchers and teachers both need a few years experience with indicators before we can see if they can help to improve instruction. As the system operates from year to year, we may see hypotheses generated from the chalk-face about causes of variations in examination effectiveness.

The Stabilities of SEIs

Table 7 shows the correlations between school effectiveness indices on four dependent variables. These were computed for three groups of students: those taking each of the indicated subjects. The cognitive variable, the residuals for *A-grades*, once again dominated such relationships as there were. As predictable, the multilevel model indicated more stability than OLS residuals, reflecting the reduction in error variance.

The table does not lend strong credibility to the notion that schools doing well with students in one aspect will be effective in all aspects. However, neither does it tell us that schools were differentially effective across multiple outcomes, because the low correlations may have been due to noise or measurement error. An approach which takes account of correlations among the dependent variables is needed and could be provided with a three level multivariate approach, as mentioned above.

Discussion

There are a number of problems with the ALIS data which are difficult to surmount in practical ways. The small sample sizes obtained even when we obtain 100 percent of those in the groups is simply a feature of A-level provision. There is also the problem that in large schools and colleges there were several classes (teaching units) within a single subject. Analyses conducted in previous years (e.g., Fitz-Gibbon, 1983) seemed to indicate that these classes-within-schools were as different as were different schools. The number of level-two units would go up considerably by using the class rather than the whole department.

If models are to be fitted to measured outcomes of schooling, then multilevel modelling would seem to be the way forward. It takes account of intra-school correlations so that its estimates are more accurate, it provides tests of slope heterogeneity (i.e., different relationships between input and outputs in different schools); it enables tests to be made for contextual effects, and it enables assessments to be made of the reliability of school effectiveness indices.

All these advantages are important for research purposes, but two major problems give pause for concern regarding the use of multilevel modelling in the kind of confidential, formative, feedback system which the A Level Information System represents. The first problem is the shrinkage which can be considerable in small samples and which would act as a kind of statistical "expectancy effect": schools with disadvantaged intakes and small samples would have their effectiveness pulled down towards that expected in the larger sample. Schools with favorable intakes and small samples would have their effectiveness indices pulled up. Secondly, the use of multilevel modelling might remove control of the information system from the users: the school personnel. As long as the residuals are based on OLS, they are easily explained. The users can see how the data are handled. The use of OLS corrections for intake is such an improvement over the current practice of publishing raw results that we must wonder if the time is right to push the analyses further.

The solution may well be to report both OLS and multilevel analyses and to continue to warn that different questions and different assumptions lead to different answers, and this complexity in the world simply has to be borne. It is our experience that once schools begin to look at real data they see that the very complexity evident in real data is some guard against facile and premature judgments. And the best defence against judgments based on poor data, such as raw examination results, is to have better data.

Modelling Or Rapid-fire Graphical Bootstrapping (RGB)?

Whilst attempts to model must continue, there may be some important alternative strategies. I had the pleasure of listening to John Tukey once

and remember his saying that there was no way you could spend too much time plotting data and looking at it. The eye is the best tool we have for recognizing patterns. Certainly many mathematicians are spending considerable amounts of time and computing power *looking* at patterns these days, and they are trying to understand fairly lawful data. Is it only social scientists who force every phenomenon into an off-the-shelf model, such as the General Linear Model?

Given the assumptions which have to be made in order to apply regression analyses of any kind, (interval scales, linearity for example) should we perhaps try to stay closer to the actual data and look at it more carefully? How might that be done in a way which gives us a sense of how much confidence to place in whatever patterns we see . . . patterns like that in Figure 1? Obviously the computer is the tool which will make that possible. What is needed is some software that *not only plots graphs but also builds into the graphs an indication of the stability or instability of the relationships represented*. Bootstrapping (Efron, 1979, 1981 as cited in Jackson, 1986) is a technique for obtaining error estimates on sample data by repeatedly sampling the data itself. The technique could be applied to graphing. The graph would be specified by the user, as for example the graph in Figure 1, and the computer program could then apply bootstrapping procedures to re-calculating the data points. For each re-calculation the screen displaying the graph could be refreshed, possibly in a different color. It would be expected that rapid-fire bootstrapping would show graphs of unstable data fluctuating wildly but graphs of robust relationships would merely shimmer.

Jack-knifing, leaving out one data-point at a time, could be used similarly. The user should be able to stop the program and examine any individual data point which has caused a major disturbance in the observed pattern. Such an empirical, descriptive rather than inferential approach might be particularly useful for feeding back yearly data to schools.

References

Bloom, B. S. (1979). *Alterable variables: The new direction in educational research.* Edinburgh: Scottish Council for Research in Education.

Carver, R. P. (1975). The Coleman report: Using inappropriately designed achievement tests. *American Educational Research Journal, 12*(1), 77-86.

Dyer, H. W. (1968). School factors and equal educational opportunity. *Harvard Educational Review,* Winter, 38-56.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1-26.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika, 68,* 589-599.

Fitz-Gibbon, C. T. (1983). *Confidential, measurement-based, self-evaluation (COMBSE Project), Reports 1 and 2.* School of Education: University of Newcastle Upon Tyne.

Fitz-Gibbon, C. T. (1985). Using audio tapes in questionnaire administration. *Research Intelligence, 19,* 8-9.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin.

Jackson, P. R. (1986). Robust methods in statistics. In Lovie (Ed.), *New developments in statistics for psychology and the social sciences.* London and New York: The British Psychological Society and Methuen.

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effects of family and schooling in America.* London: Allen Lane.

Kelly, A. & Smail, B. (1984). Sex differences in science and technology among eleven-year-old schoolchildren. *Research in Science and Technological Education, 2*(2), 87-106.

Murnane, R. J. (1987). Improving education indicators and economic indicators: the same problems? *Educational Evaluation and Policy Analysis, 9*(2), 101-116.

Rasbash, J., Prosser, R., & Goldstein, H. (1988). *ML2 software for two-level analysis.* London Institute of Education, London University.

Raudenbush, S. (1989a). "Centering" predictors in multilevel analysis: Choices and consequences. *Multilevel Modelling Newsletter, 1*(2), 10-12.

Raudenbush, S. (1989b). *Personal communication.*

Smith, D. J., & Tomlinson, S. (1989). *The school effect. A study of multi-racial comprehensive schools.* London: Policy Studies Institute.

Willms, J. D. (1985). Catholic school effects on academic achievement: New evidence from the High School and Beyond follow-up study. *Sociology of Education, 58,* 98-114.

Wittrock, M. C., Marks, C., & Doctorow, M. (1978). Generative processes in reading comprehension. *Journal of Educational Psychology, 70*(2), 109-118.