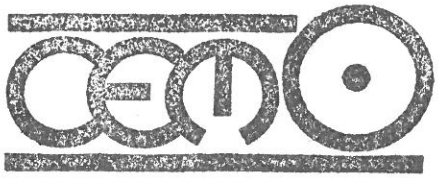

PUBLICATION NO. 32

THE STABILITY OF SCHOOL EFFECTIVENESS INDICATORS

P. B. Tymms and C. T. Fitz-Gibbon

Moray House, Edinburgh and
CEM, School of Education
University of Newcastle upon Tyne

1990

Abstract

The reliabilities of school and departmental cognitive and attitudinal educational indicators were examined at A level and found to be generally very satisfactory although some problems were noted in the measurement of the attitudes to optional subjects.

Some connection was found between the exam indicators in different subjects within a school but the correlations were generally small and even smaller for attitude indicators. On this basis the concept of a whole school effect is questioned.

Measures of stability of indicators over the years were examined, but found to vary according to the indicator, some varying markedly with others remaining fairly constant. It also proved possible to identify, with some clarity, the degree to which departments had changed over a two year period, giving more confidence in the measurement procedure and emphasising the need for specific qualitative data about schools which could help to explain the change.

Introduction

This paper is based on an analysis of the 1989 survey of the A Level Information System (ALIS) (Fitz-Gibbon, 1985, 1990a, b and c, Fitz-Gibbon et al, 1989). The data used were the students' A level grades in the subject, measurement of attitudes to the subjects as well as to school, prior achievement and gender. The A level grades were examined separately but they were also averaged so that the mean grade could be used as a variable in its own right.

Each of the outcome variables were first examined for its reliability as a measure of educational units having controlled for various combinations of input measures. Then the relationship between the various measures of schools and sub-units in the schools in 1989 were considered. Finally, the stability of a particular measure over the year was considered as well as the correlation between one type of measure in 1989 with a different type of measure in 1990.

From these examinations of reliabilities and correlations conclusions were reached regarding the appropriate units of analysis for performance indicators, and suggestions were given for future research.

RELIABILITY OF SCHOOL EFFECTIVENESS MEASURES

Mandeville and Anderson (1987) looked at the reliability of maths and reading measures in grades 1 to 4 for 423 schools. They used school level data and ordinary least squares (OLS) regression to control for prior achievement test scores. Schools with less than 20 students were omitted from the analysis and reliabilities were estimated by comparing the effectiveness indicators (EIs) of two random subsamples and then stepping up the correlations to estimate the value for the total sample. For reading the median was .78 and for mathematics it was .86.

Dyer et al (1969) also worked with two random subsamples and OLS regression and stepped up the correlations to give a median value of .83 for eighth grade students across a range of tests.

Marco (1974) reported a reliability of .83 for reading scores using school level data and OLS regression controlling for prior test scores.

More recently it has been possible to estimate the reliabilities of EIs using multilevel software (Aitkin and Longford, 1986, Goldstein, 1988, Raudenbush and Bryk, 1986). In particular the formula given by Goldstein (1987), for the 'shrinkage factor' provides direct access to the reliability of a particular EI based on the number of students in a particular sample. The formula is:

$$n_j\sigma_u^2(n_j\sigma_u^2 + \sigma^2)^{-1}$$

where n_j = number of students in a school

σ_u^2 = school level variance

σ^2 = pupil level variance.

Table 1 gives details of the multilevel analysis applied to the 1989 ALIS data for 10 A level subjects and for students' mean grades on the subjects taken.

Reliabilities were calculated for departmental sizes of 12, 50 and the mean departmental size.

Insert table 1 about here

Insert diagram 1 about here

Diagram 1 shows how the reliability of the school level residuals (School Effectiveness Indicators EIs) varies with sample size for A level mathematics. It is apparent that the reliability rises rapidly with sample size and that with n>50 it is very satisfactory (cf Raudenbush and Bryk 1986) and that even with n=12 the EIs are worthy of consideration.

The reliabilities varied a little from subject to subject but it is not possible to make a general statement about their variation by curriculum areas, as was suggested by Mandeville and Anderson (1987). The lowest of the reliabilities was for General Studies which in many respects may be regarded as a high level ability test (Forrest et al, 1970) and could therefore be less susceptible to teaching effects. The average class contact time per week for General Studies was 1.5 hours compared with 5 hours for all the other A level subjects on average. The highest reliability was recorded for Economics.

A surprising aspect of the reliability measures for the EIs is that they do not apparently reflect the reliabilities of the individual A level grades awarded by the Boards. Data was not specifically available for this but one might suppose that the A level grades in maths and science would be be

accompanied with less error at the student level than those in the humanities. The reason for this is the format of the examinations. For example the Joint Matriculation Board History mark was heavily influenced by essay work and if there was a tendency for markers to avoid very high and very low marks, then the final discrimination must be low. Indeed the JMB examiners' reports for History 1989 confirm that for Syllabus A, for candidates taking the exam entirely through written papers, the difference between grade boundaries was typically 9 marks out of 200. On the other hand in JMB Chemistry there was a multiple choice section and many short answer type questions and a further two questions which have sub-parts to them. The 1989 JMB exam report for Syllabus A Chemistry showed that typically 22 marks out of 340 separated grade boundaries.The greater discrimination of Chemistry compared with History at the student level was further evidenced by the correlations with prior achievements scores which were .66 and .48 respectively.

But the reliability estimates depend not only on variation at the student level but also on the variation at the department level, a point made by Smith and Tomlinson (1989). Thus, in the case of A levels, it could be that there was more accurate measurement of student performance in say Physics compared to History but that Physics departments were more homogeneous than History departments and so the reliabilities of the two effectiveness indicators were similar.for similar department sizes

The smaller reliabilities, recorded for equal sample sizes, for the exam mean compared to individual subjects are due to the smaller variation between schools than between departments for individual subjects. However, since the mean score measure can be based on a full year group rather than just a department the reliability calculated for the mean sample size is higher than all the other individual subjects with the exception of Economics.

The reliabilities were also calculated for class-based as opposed to department-based indicators. Unfortunately a third level (school) could not be included in the model since in many schools the class was the department. The data are presented in Table 2. In every case the reliabilities were higher for class measures than they were for department measures, for the same sample sizes, although there is a price to pay since classes are generally smaller than departments but when the reliabilities for mean sizes are considered there was little difference between departmental and class reliabilities.

Insert table 2 about here

The process was repeated for attitude measures at the department level only and the data are presented in Table 3. It is apparent that the attitude measures generally produced reliabilities considerably lower than exam scores although General Studies was an exception to this. It could be that producing EIs for attitudes to subjects is particularly problematic when the subjects have been specifically chosen by students, who presumably had a clear view when option choices were made. This would explain the high reliability for General Studies which is often not an optional subject. The measures for Physics and Biology were high compared to other subjects but for some subjects the reliabilities were so low as to make their use debatable. Just why this should be so is not clear. This difference between Chemistry and Physics is particularly puzzling.

Insert table 3 about here

The attitudes were all measured with the same six questions on 5 point Likert-type questions and had internal reliabilities (Cronbach's alpha) of around .8.

The attitude to school scale (ATTSCH) was also measured using six questions which were answered on 5 point Likert-type scales with alpha equal to .81. Its reliability was very satisfactory especially when it is remembered that it can be used to report a whole school EI rather a departmental indicator. In fact, it had the highest reliability of any EI in the study, for the number of students available. Nevertheless, it may generally be stated that indicators were more reliable when used to report academic attainment than attitudes.

EFFECTIVENESS INDICATORS COMPARED WITHIN ONE YEAR

Reading and Maths EIs were compared within grades by Mandeville and Anderson (1987). They found correlations varying from .6 to .7. Whereas Helmstadter and Walton (1985) found EIs for Reading, maths and language to vary from .7 to .9 for third to sixth graders.

Willms and Raudenbush (1988) suggested that the rather low stabilities of EIs were due to sampling and measurement errors on the one hand and a failure to distinguish between types of EI on the other. They identify Type A and Type B effects. Type A effects are calculated by taking into account student level explanatory variables only but Type B effects result from including in the model contextual influences. They calculated Type A and Type B EIs for English, arithmetic and overall attainment in the Scottish Certificate of Education for 20 secondary schools for two separate years. They found correlations between the Type A effects of between .4 and .7 but for Type B effects of between .2 and .7.

In the data set under consideration no contextual effect was found (except for French exams and Economics attitudes) during the modelling and so all EIs, with these two exceptions, are in a sense Type A EIs. However, since contextual effects were tried in the model and found not to be important one could argue that Type A and Type B effects are synonymous here, although the possibility of an important contextual variable on which there was no data is certainly conceivable. Table 4 gives the correlations for OLS and ML exam residuals for ten A level exams in 1989 in the ALIS project between the two types of measure. The OLS and ML measures where very similar, correlating highly with one another; a not unexpected result since discrepancies become most apparent if the intra school correlations are high or if contextual effects are important. The ML values may be expected to be higher than the OLS values since they are not distorted by the kind of exceptional results which can appear with small samples, although the ML correlations could be biased since the 'shrinkage' will tend to give similar values to small departments which tend to be clustered in certain sixth forms. (Efron and Morris 1975) However, the mean ML and OLS correlations were almost identical at about .15.

Insert table 4 about here

Insert tables 5-8 about here

However, the most efficient way to estimate the correlation between two EIs is to use a model in which scores are nested within individuals which are

nested within schools. The size of the data set and the nature of the PC version of the software was such that only two scores could be modelled together at any one time. The process was therefore only tried on a subset of all possible pairs. The correlations for these analyses are shown in Table 5. There was some tendency for these correlations to be higher than either the OLS or the ML residual correlations, but not dramatically so. The correlations for Biology with Physics and with Chemistry in particular were high. Tables 6 and 7 give the numbers of students that were entered for both exams and the numbers of schools involved. Table 8 gives the covariance with standard errors on which the correlations were calculated. It is clear that there were fairly considerable errors associated with the correlation estimates and that the resultant correlations were not very different, in view of the associated errors, from the ML and the OLS residual correlations. The two largest rises were associated with the two smallest numbers of overlapping students.

For this subset, the correlations varied from .46 to .15 with mean value of .32. These correlations are low, and yet table 4 indicates that they were amongst the highest of the between subject correlations, and this would seem to imply that there is little sense in which a whole school exam indicator can be formulated. Presumably, departments are semi-autonomous and can be independently successful (a 'nearly decomposable' system (Simon, 1969)). And yet there seems to be a great desire among researchers, inspectors and heads to look for whole school effects and to have single numbers to identify effective schools. It has already been noted above that a reliable EI for the exam mean for a school can be estimated but that the proportion of variance accounted for at the school level is somewhat lower than the proportion of variance accounted for at the department level.

An alternative way to estimate a school effect would be to amalgamate the individual department ML residuals. When this is done a scale is formed with an internal consistency (Crobach's alpha) of .53 (Economics and General Studies were excluded because there were several institutions which did not enter candidates for those subjects) giving no great confidence in the overall scale. It might perhaps be thought that there would be greater consistency within say the science/maths department as a whole but the reliability for such a scale was .57: Again rather low.

The correlations for ML attitude residuals are presented in Table 8. It has already been noted that the reliabilities of the attitude EIs were rather low, but it is nevertheless sobering to note that the mean correlations between subjects was .01! Indeed an examination of the subject/subject figures of Table 8 lead to the conclusion that even in those subjects with the highest reliabilities (Physics, Biology and General Studies) there was only the slightest evidence of consistency within schools.

The mean correlation between ATTSCH and subject attitudes was .14, again a very low figure especially when one considers that the modelling of the attitude measures outlined in table 3 only took account of cognitive and gender measures and not prior attitudes. There may have been individuals whose disposition was such that they tend to answer attitude questions positively and that therefore there should be positive correlations between all attitude measures. But perhaps the distribution of such individuals between schools was largely without bias.

Finally the correlations between the ML attitude residuals and the ML exam residuals are presented in table 9. Once again the correlations were low between exam and attitude EIs. In the same subject the average was .16, although the most reliably measured subject attitude (General Studies) did

correlate .39 with the corresponding exam EI, however, this was coincidentally based on the lowest number of schools (42).

Insert table 9 about here

Generally the connection between the attitudes in one subject and the exam results in another was non-existent (mean r=-.01) but there was a very slight tendency for those schools which scored highly on the attitude to school measure to have better than expected exam results (mean r=.14)

THE STABILITY OF SCHOOLS EFFECTS ACROSS YEARS

When looking at the correlations between grades Mandeville and Anderson (1987) found a median of .06 for reading and a median of .13 for mathematics and referred to the values as 'discouragingly small'. Forsyth (1973) looked at two successive years of 12th grades on standardised test scores and found correlations amongst residuals to have a median of .28. Mathews, Soder, Ramey and Sanders (1981) found correlations for successive years for the same subject areas from -.24 to .44.

Willms and Raudenbush (1988) consider the stability of results from year to year in some detail and provide an exemplar of the way in which to proceed with the analysis, distinguishing as was mentioned above between Type A and Type B effects as well as between the stable and unstable components of EIs, which they were able to estimate separately in their analysis, which uses multilevel software.

The reliability of estimates of individual EIs for one measure for one year have already been covered but it is important to note that when dealing with longitudinal data it is possible to estimate the reliabilities of the stable and unstable components of the EIs and to do this for Type A and Type B effects. Willm and Raudenbush's estimates for the reliabilities for 20 schools varied from .78 to .36 for stable component and from .43 to .21 for the unstable component. They also reported pre post correlations for which the "true" correlations varied from .87 to -.01 for Type A and Type B effects between between 1980 and 1984.

ATTITUDE TO SCHOOL OVER TWO YEARS

For the data under consideration Attitude to School, Maths and English exams EIs were selected for more detailed consideration, since they were reliable indicators based on the largest numbers of schools and students.

Following the suggestions of Willms and Raudenbush (1988) in their example for T=2 the year was coded -.5 for 1988 and .5 for 1989. The results of the modelling are given in Table 10. From these it may be seen that both the stable component of the ATTSCH measure and the unstable component (ie the change over the two years) vary across schools and would appear to be measured with reasonable reliability.

Insert table 10 about here

The best estimate of the correlation between the school EIs for the two years may be given by the formula given by Willms and Raudenbush (1988) for the "true" correlation. It was .74, .77 and .14 for Attitude to School, Maths exams and English exams respectively. This first is a high figure and it is worth

speculating on the reasons for this. The questions on which the attitude scale were based asked amongst other things if they would advise others to take their A levels at that institution. Now a question such as this asks not simply for a student's own perception but for an overall assessment of a college or school and a student must in such a situation, to some extent, take account of what others have said about other parts of the institution, removed from first hand knowledge. In other words the scale will tap the rumours about, and reputation of, an institution which might be more stable than harder more direct measures of a department or school. But of course an institution's reputation is part of a school and a dynamic must exist between hard data and hype.

The apparently high stability of the Maths EI is in stark contrast to that of English and it could have something to do with the precise nature of a Maths exam compared to an English exam. However, that would seem to be unlikely in view of the reliabilities in Table I. It may also be related to the syllabuses of English Literature which involve changes in set books regularly as opposed to the consistency of Maths. But whatever explanations are advanced for the different stabilities, more data than is available would be required to distinguish between them and any thorough explanation would have to take account of the opposite finding by Willms and Raudenbush (1988) for Arithmetic and English.

ENGLISH AND MATHS OVER SEVEN YEARS

The main data within ALIS relates to 1988 and 1989 but 10 institutions have been part of the project since 1983 and their data were available for English and Maths A levels. These data for exams only, were examined to look for stability over the years. (Fitz-Gibbon, (1989) examined the same data up to 1987 using OLS). It is, unfortunately from the measurement point of view, often the case at A level that there are small classes and this was certainly true for the

schools in question, but it was nevertheless felt to be worth examining the data since decisions in the real world sometimes have to be taken on the basis of such small numbers.

In the first instance students were modelled within years and years within schools, but with seven years and ten schools, the optimistic strategy was soon dropped in the face of modelling difficulties. It was more successful to model students within schools and to use dummies for years where appropriate. Table 11 gives the results of these models. At first sight the models would seem to be successful with moderate reliabilities for the EIs but these were based on school level variances which had errors which were larger than half the values which were estimated. This implies that there was no statistically significant difference between the overall EIs of the schools when their data were aggregated over 7 years.

Insert table 11 about here

Table 12 shows the same data remodelled with school/years at level 2. That is to say, each school's results each year was treated as a separate unit. Now the variances at level 2 were higher and large enough compared to the errors to indicate that there were indeed measurable differences between school/years. However, the reliabilities for the school/year EIs were down to .5.

These models, reported in Tables 11 and 12, imply that, when averaged over the years schools tend to have similar examination success given the kinds of pupils that they have on their courses, but that particular year groups of students in particular subjects can do noticeably better or worse than would be expected.

Insert table 12 about here

The residuals for each school were plotted by year for English and maths: the plots are shown in diagrams 2 and 3. Lines are indicated only for three schools so as not to clutter the diagrams.

Insert diagrams 2 + 3
about here

Stability is noticeable mainly though its absence in the two diagrams, but with reliabilities of .5 only exceptional cases could be expected to show any uniformity.

For English, school B appears to have a consistently 'good' record with only a -.1 in '85 to mar the positive residuals. School H, on the other hand was slightly below expected every year except '89. School C appeared to have a particularly worrying year, in '86 straddled by two relatively very successful years.

For Maths school B was consistently 'poor' except for a score of .1 in '85. Almost a mirror image of the English department in the same school. School C always appeared above or equal to the expected grade.

Both sets of EIs which appear in diagrams 2 and 3 were examined for the extent to which the figures represented reliable indicators. For English the figure was .5 and for maths .6. (Oddly similar in view of the correlation found over just two years,)

All in all the two diagrams did indicate some measure of stability but not enough to be able to pick out a consistently outstanding department and certainly not a consistent school.

DISCUSSION

The data presented in this paper has generally confirmed the finding that it is indeed possible to measure the performance of a school, or department with a reasonable degree of reliability for one year. This measurement becomes much more reliable as the number of students on which it is calculated increases, a point which has implications for the level of aggregation at which results should be reported. The data also support the almost self-evident proposition that the closer a measurement is to classroom activity then the more it is related to outcomes. In statistical terms this means that the greatest proportion of variance will be explained by the teacher, then by class membership then by department then by schools and least by LEA. But of course as the measure gets closer to a pupil's learning environment, then the smaller will be the sample on which the EI is based, So that whole school EIs might be expected to be reliably measured but to have less direct meaning for membership of the school for a particular pupil. Of course this can be viewed in two ways and whilst a whole school effect is likely to mean little for an individual pupil it will cumulatively mean a lot since the effect is distributed so widely. (Bosker 1989) argues convincingly that "though school effects might be moderate or low in terms of percentage of variance accounted for they may still have quite significant societal consequences." A balance needs to be drawn between the reliability of a EI and the degree to which it represents educational meaning for a pupil. This balance might vary from system to system and for measure to measure. Within ALIS many indicators are reported, but for those

examined in this paper, a reasonable compromise would seem to be to employ exam indicators at the department level as has been done since 1983 . Smith and Tomlinson (1989) also recommend. that the department be considered the unit of analysis. The school attitude measure would be most appropriately reported at the school level of aggregation

The data presented also indicate that attitude measurement is not easy in the sense that in order to obtain reliable measures large numbers of students must be consulted on a systematic basis. For A levels the only really reliable measure was for a whole school, although with sample sizes over 50 in some departments the reliabilities were generally good. This has implications for school inspection by individuals or groups of individuals.

As regards the connection between EIs within the same year, the data indicate that it is possible to measure different parts of an education institution and to find different results. In particular, departments may be given effectiveness indicators which are largely independent of similar indicators of other departments. Furthermore, departments can be given indicators of academic success as well as student attitudes to the institution and these measures are nearly orthogonal.

Witte and Walsh (1990) give theoretical and practical reasons for the school often being treated as the unit of analysis by effective schools research but the empirical evidence shows a more complex picture as it is examined more closely.

Stability over the years is certainly present to some extent, particularly for A-level Maths and attitudes to school, but the reasons for the stability or lack of stability (for English) may only be hypothesized. Intriguingly, the models

presented in this paper indicate that it is possible to pin-point quantitatively, those departments where the effectiveness indicators have changed and the degree of change. Put differently, the changes in EI from year to year would appear not to be entirely due to errors of measurement: there are genuine changes in effectiveness. Several hypotheses have been noted in this paper but more research is needed to try to tease out reasons for the changes. The ALIS database does contain extensive data which could be examined but qualitative, longitudinal data is really needed before firmer theories can be formulated.

REFERENCES

Aitkin, M. and Longford, N. (1986) Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A.* 149(1) pp 1-43

Bosker, R.J. & Scheerens, J. (1989) Criterion-Definition, Effect Size and Stability, Three Fundamental Questions in School Effectiveness Research in Creemers, B. Peters, T & Reynolds, D (ED) *School Effectiveness and School Improvement* Swets & Zeitlinger

Efron B. and Morris C. (1975) Data analysis using Stein's estimator and its generalisations *Journal of the American Statistical Association* 74(4) 396-430

Fitz-Gibbon, C.T. (1985) A-level results in comprehensive schools: The Combse project, year 1. *Oxford Review of Education.* 11(1) pp 43-58

Fitz-Gibbon, C.T., Tymms, P.B. and Hazelwood, R. D. (1989) Performance Indicators and Information Systems. In Reynolds, D. Creemers, B.P.M. and Peters, T (Eds) *School Effectiveness and Improvement* Netherlands: RION Institute for Educational Research and University College of Cardiff. ISBN 1 872330 00 2

Fitz-Gibbon, C.T. (a) (1990) An Up-and-Running Indicator System. In Fitz-Gibbon, C.T. (Ed.) *Performance Indicators: a BERA Dialogue* Clevedon, Avon: Multilingual Matters pp 88-95.

Fitz-Gibbon, C.T. (1990) (b) Performance Indicators: educational considerations. In Levacic, R. (Ed) *Financial Management in Education* Milton Keynes:Open University Press

Fitz-Gibbon, C.T. (1990) (c) Multilevel modelling in an indicator system. In Raudenbush S.W. and Willms, J.D. *Pupils, classrooms and schools: international studies of schooling from a multilevel perspective.* London and New York? : Academic Press pp 45-61

Forrest, G.M., Smith, G.A. and Brown,M.H. (1970) *General Studies (Advanced) and academic aptitude.* Manchester: Joint Matriculation Board.

Forsyth R. A. (1973) Some empirical results related to the stability of performnce indicators in Dyer's stedent change model of an educational system. *Journal of Educational Measurement* 10(1) 7-12

Goldstein, H. (1987) *Multi-level models in Educational and Social Research.* London:Griffin

Helmstadter G. C. and Walton M. A. (1985) The generalizability of residual indexes of effective schooling *AERA paper* Chicargo

Mandeville, G.K. and Anderson, L.W. (1987) The stability of school effectiveness indices across grade levels and subject areas. it Journal of Educational Measurement 24(3) pp 203-216

Marco, G.L. (1974) A comparison of selected school effectiveness measures based on longitudinal data. *Journal of Educational Measurement* . 11(4) pp. 225-234

Mathews T.A. Soder J. B. Ramey M. C. and Sanders G. H. (1981) Use of district test scores to compare the academic effectiveness of schools. *AERA paper* Los Angeles

Rasbash, J., Prosser,R. and Goldstein, H. (1988) *ML2 Software for Two-level analysis.* London:Institute of Education, London University

Raudenbush,S. and Bryk,A.S. (1986) A hierarchical model for studying school effects. *Sociology of Education* 59, pp 1-17

Simon, H.A. (1969) *The Sciences of the Artificial* Boston: The MIT Press

Smith, D.J. and Tomlinson, S. (1989) *The School Effect. A study of multi-racial comprehensives. London:Policy Studies Institute./SE*

Tymms, P. B. and Fitz-Gibbon, C.T. (1990) A comparison of exam boards: 'A' levels. *Oxford Review of Education* in press

Williamson, J. and Fitz-Gibbon, C.T. (1990) On the lack of impact of information. *Educational Management and Administration* . 18(1) pp37-45

Willms, J.D. & Raudenbush, S.W. (1988) Estimating the stability of school effects with a longitudinal, hierarchical linear model. it AERA-paper New Orleans

Witte, J.F & Walsh, D.J. (1990) A Systematic Test of the Effective Schools Model. *Educational Evaluation and Policy Analysis* Vol 12 No 2 pp 188-212

Diagram 1

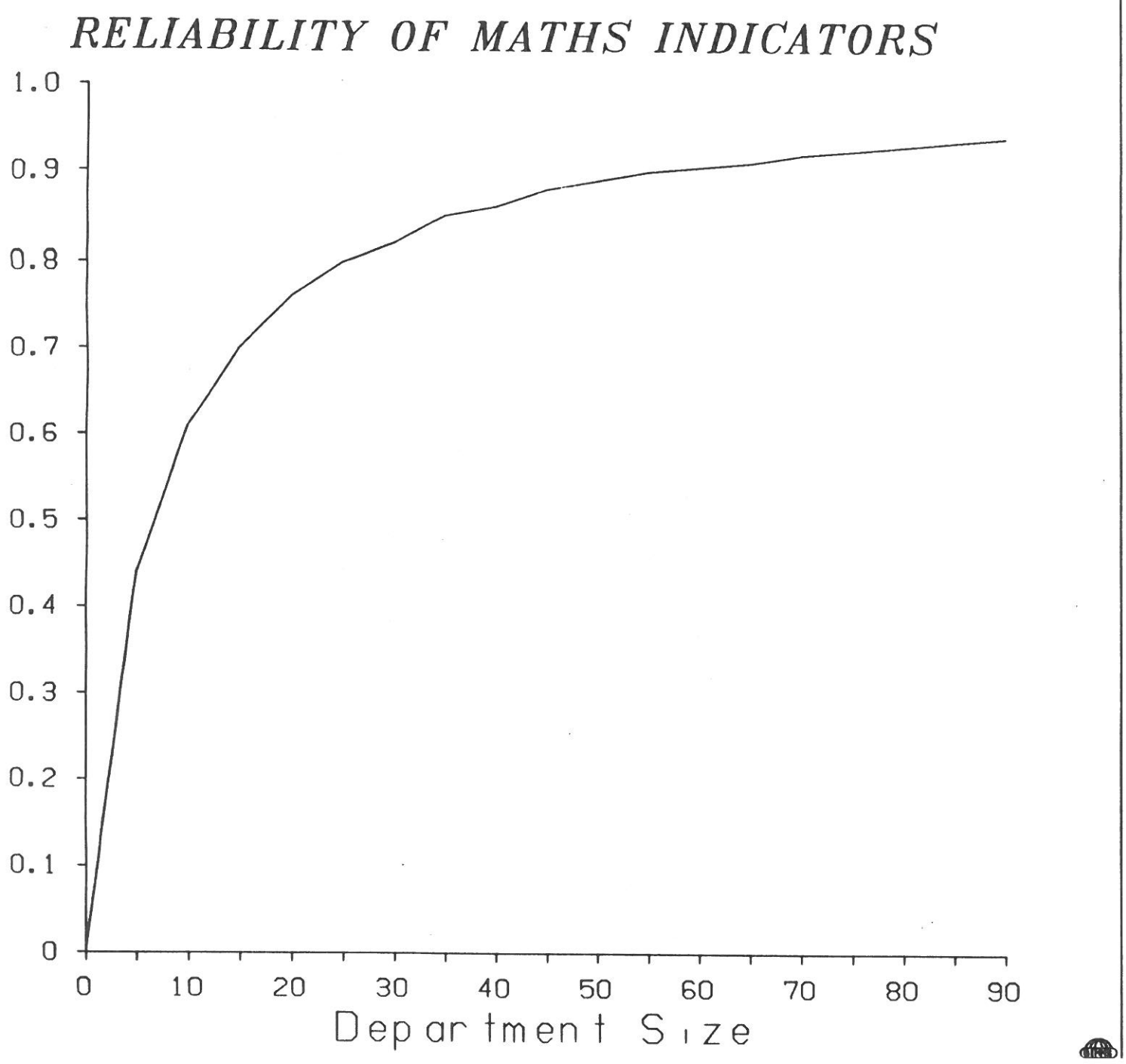
RELIABILITY OF MATHS INDICATORS
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
Department Size

DIAGRAM 2

ENGLISH

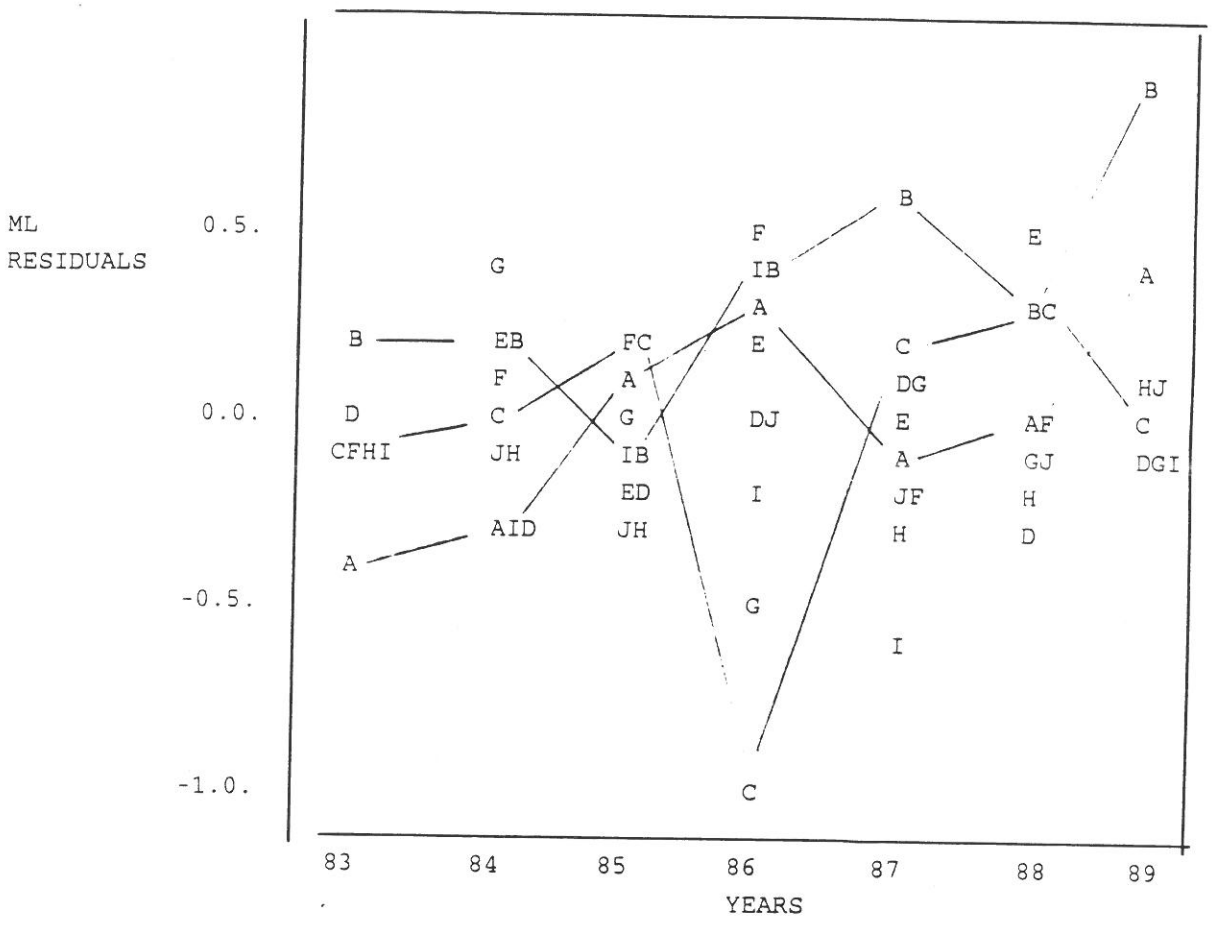

DIAGRAM 3
MATHS

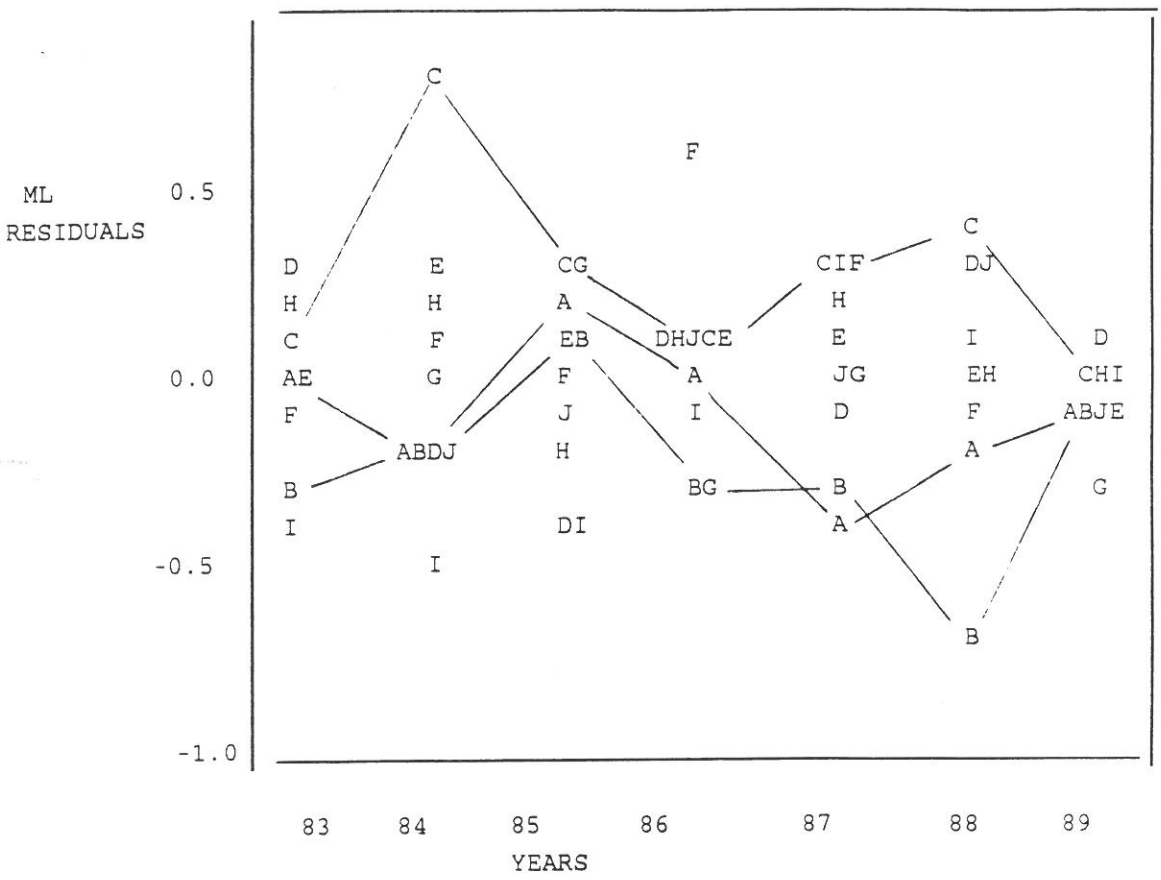

TABLE 2 Class Level

	MATHS	PHYS	CHEM	BIOL	ECON	GEN ST	ENG	FREN	HIST	GEOG
/ariance School level)	.56	.30	.34	.25	.62	.25	.36	.39	.46	.37
/ariance Student level)	2.33	1.69	1.75	1.81	1.91	2.49	1.82	1.73	2.01	1.88
ıtra class correlation	.19	.15	.16	.12	.25	.09	.17	.18	.19	.16
Mean n within class	12.8	10.4	10.5	8.2	10.4	28.4	12.5	6.2	9.6	9.7
Reliability :-										
n = 12	.74	.68	.70	.62	.80	.55	.70	.73	.73	.70
n = 50	.92	.90	.93	.87	.94	.83	.91	.92	.92	.91
n = Class mean	.75	.68	.67	.53	.77	.74	.71	.58	.69	.66
N = Classes	110	85	82	87	65	41	81	69	80	73

Table 2 was based on the same models as Table 1 but with classes rather than schools at level 2.

TABLE 1 Exam

	Maths	Phys	Chem	Biol	Econ	Gen S	Eng	Fren	Hist	Geog	MEAN GRADE
FIXED											
CONS	-7.51	-9.72	-8.81	-7.20	-5.95	-6.31	-5.03	-8.21	-3.43	-5.78	-6.11
AVOG	1.66	2.04	1.93	1.74	1.45	1.51	1.35	1.63	1.02	1.51	1.46
SEX BOYS=0 GIRLS=1	- .39	- .62	- .48	- .44	- .50	- .78	ns	- .44	- 2.3	- .41	- .27
% FEMALE	ns	ns	ns	ns	ns	ns	ns	1.33	ns	ns	ns
MEAN AVOG	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
SEX x AVOG	ns	ns	ns	ns	ns	ns	ns	ns	.38	ns	ns
VARIANCE (School level)	.38	.26	.21	.23	.61	.22	.28	.34	.33	.35	.12
VARIANCE Student level)	2.43	1.72	1.83	1.81	1.93	2.53	1.80	1.82	2.09	1.86	1.63
Intra School Correlation	.14	.13	.10	.11	.24	.08	.13	.16	.14	.16	.07
Reliabilities: n = 12	.65	.64	.58	.60	.78	.51	.65	.69	.65	.69	.47
n -=50	.89	.88	.83	.86	.94	.81	.89	.90	.89	.90	.79
n=school mean	.77	.67	.61	.58	.81	.72	.71	.57	.66	.68	.81
Variance explained at School level	21%	47%	43%	57%	8%	22%	1.7%	32%	13%	14%	37%
Variance explained at Pupil level	33%	49%	46%	40%	38%	32%	34%	38%	26%	32%	40%
Mean n within schools	22.0	13.4	13.4	11.3	13.3	29.4	15.7	7.1	12.5	11.2	57.1
N = school	68	69	67	67	54	42	69	63	67	66	70

NB In every case the intercept alone was allowed to vary across schools and the coefficients for AVOG and SEX were fixed but only after examining alternative models

ns = non-significant

TABLE 4

OLS EXAMS

ML EXAMS	MATH	PHYS	CHEM	BIOL	ECON	G STD	ENGL	FREN	HIST	GEOG
MATH	.97	.25	.20	.25	.45	.36	.16	-.09	.01	.11
PHYS	.25	.93	.17	.35	.14	.16	.33	.18	-.02	.15
CHEM	.28	.30	.68	.22	.14	.16	.29	.20	-.19	.33
BIOL	.27	.27	.08	.91	.23	.28	.10	.13	-.15	.08
ECON	.44	.13	.10	.23	.93	.55	.08	.01	.14	.14
G STD	.24	.27	.05	.14	.49	.77	.31	.14	.12	.04
ENGL	.22	.39	.28	.11	.08	.45	.92	.12	.03	.08
FREN	.06	.22	.15	.20	.03	.15	.22	.86	.13	.20
HIST	.00	-.12	-.21	.05	.12	.01	-.07	.03	.89	.02
GEOG	.08	.18	.35	.12	.09	.08	.16	.12	-.17	.91

OLS with ML on diagonal
Mean OLS correlation = .16
Mean ML correlation = .15
Mean OLS/ML diagonal correlation = .88

TABLE 3 ATTITUDES

	Maths	Phys	Chem	Biol	Econ	Gen S	Eng	Fren	Hist	Geog	Att Sch
FIXED											
CONS	1.69	1.57	1.93	1.72	4.71	2.86	3.63	1.74	2.05	2.54	3.18
AVOG	.23	.22	.16	.26	.22	-.04	.17	.211	.19	.11	.083
SEX BOYS=0GIRLS=1	ns	-.14	-.11	ns	-.11	-.14	-.26	ns	ns	-.12	.087
% FEMALE	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
MEAN AVOG	ns	ns	ns	ns	-.57	ns	ns	ns	ns	ns	ns
SEX x AVOG	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
VARIANCE											
(School level)	.021	.06	.014	.056	.021	.082	.016	.037	.038	.032	.050
(Student level)	.64	.55	.62	.51	.62	.49	.58	.73	.60	.60	.43
Intra School Correlation	.02	.10	.02	.10	.03	.14	.03	.05	.06	.05	.10
Relibilities:											
n = 12	.28	.57	.21	.57	.29	.67	.24	.38	.43	.23	.58
n = 50	.62	.85	.53	.85	.63	.89	.58	.71	.76	.73	.85
n=school mean	.42	.60	.23	.55	.31	.83	.30	.26	.44	.38	.87
Variance explained at School level	-10%	-2%	6%	14%	30%	0%	-1%	-19%	-34%	-6%	0%
Variance explained at Student level	4%	4%	2%	5%	4%	1%	2%	3%	3%	1%	1%

NB In every case the intersept done was allowed to vary across schools and the coefficients for AVOG and SEX were fixed after examining alternative models

ns = non-significant

TABLE 5
SEI exam correlations
(Subjects within pupils within schools)

	Math	Phys	Chem
Phys	.21		
Chem	.30	.38	
Biol	.48	.43	.15

TABLE 6
Numbers of students in common for table 5

	Math	Phys	Chem
Phys	739		
Chem	627	490	
Biol	289	133	401

TABLE 7
Numbers of schools in common for table 5

	Math	Phys	Chem
Phys	68		
Chem	66	65	
Biol	64	62	61

TABLE 8
Covariance for calculation of table 5 data

	Math	Phys	Chem
Phys	.07 (.06)		
Chem	.09 (.06)	.09 (.05)	
Biol	.13 (.06)	.11 (.06)	.03 (.05)

TABLE 8
ML Attitudes

	MATH	PHYS	CHEM	BIOL	ECON	G STD	ENGL	FREN	HIST	GEOG
PHYS	.03									
CHEM	-.35	.17								
BIOL	-.04	.20	.05							
ECON	.15	.11	-.23	-.08						
G STD	-.07	-.21	-.17	-.06	.03					
ENGL	-.08	-.04	.16	.14	.01	.13				
FREN	.07	.07	.15	-.03	-.07	.15	-.07			
HIST	-.20	-.10	-.16	.07	-.08	.12	.08	-.10		
GEOG	-.14	.18	-.11	.42	.19	.08	-.06	.02	.03	
SCHOOL	-.06	.23	.16	.44	.17	-.08	.23	-.17	.16	.29

MEAN SUBJECT/SUBJECT ATTITUDE = .01
MEAN SUBJECT/SCHOOL ATTITUDE = .14

TABLE 9

ML Exams

		MATH	PHYS	CHEM	BIOL	ECON	G STD	ENGL	FREN	HIST	GEOG
	MATH	.25									
	PHYS	-.17	.21								
	CHEM	.07	.18	.21							
ML	BIOL	.24	.22	.03	.38						
ATTITUDES	ECON	-.15	-.21	-.11	-.01	-.11					
	G STD	-.18	.12	-.36	-.20	.24	.39				
	ENGL	.10	.14	.04	.05	.05	-.10	-.12			
	FREN	-.10	.29	.19	-.08	-.20	-.14	.30	.05		
	HIST	-.10	-.04	-.21	.10	-.15	.01	-.04	.09	.05	
	GEOG	-.08	.05	-.05	-.09	-.10	-.12	-.01	.13	.04	.31
	SCHOOL	.21	.11	.06	.18	.06	.17	.18	.24	.00	.21

Mean exam/attitude same subject = .16

Mean exam/attitude different subject = -.01

Mean exam/school attitude = .14

TABLE 10

Attitude to School 1988-89

Fixed	
CONS	3.18 (.08)
AVOG	.082 (.013)
SEX	.075 (.018)
YEAR (1988 = -.5) (1989 = +.5)	.046 (.03)
Random	
Variance (CONS)	.0336 (.0076)
Covariance	.0102 (.0059)
Variance (year)	.0225 (009)
Variance (Student level)	.427 (.008)
n = schol mean	79
n = school	70
"True" correlation 1988/89	.74

TABLE 11

with school at level 2

1983 - 1889

	MATHS	ENGLISH
fixed		
CONS	-8.31 (.50)	-6.19 (.45)
AVOG	1.77 (.09)	1.50 (.08)
SEX	- .24 (.12)	ns
random		
Variance (school level)	.071 (.048)	.072 (.051)
Variance (pupil level)	2.38 (.12)	2.24 (.12)
Reliability	.71	.66

Alternative models with dummies for years were examined but found wanting.

TABLE 12

wiht (school/year) at level 2

1983 - 1989

	MATHS	ENGLISH
fixed		
CONS	-8.17 (.49)	-6.29 (.44)
AVOG	1.76 (.09)	1.52 (.08)
SEX	- .65 (.24)	ns
'88 (dummy)	-.65 (.24)	ns
random		
Variance (school/year level)	.155 (.062)	.207 (.049)
Variance (pupil level)	2.26 (.12)	2.10 (.12)
n = school/year	66	64
n = school/year mean	12.4	10.3
Reliability	.46	.50

Alternative models with dummies for years were examined but found wanting.