From academe to classroom: Translating the results of multilevel modeling into changed practice

Dr John DeCourcy
Macquarie University & St Andrews College Marayong

Paper/workshop presented at the
ACSPRI Social Science Methodology Conference
The University of Sydney, 10-13 December 2006

Multilevel modeling is suited to the analysis of relationships within nested sets of data such as those found in student achievement within classes within schools within systems. However the interpretation of findings from fitted multilevel models can be difficult for practitioners not trained in the techniques of multilevel analysis. Further, the different audiences for the research (e.g., classroom teachers, principals, system administrators) ask quite different questions of the data. This paper will: (1) outline the multilevel methodology used over six years for approximately 125 secondary schools in providing a learning-gain analysis; (2) comment on whole-of-system findings and changes observed over the years that schools have become engaged with the analysis; and (3) demonstrate some of the means whereby the findings from apparently complex statistics of MLwiN can be interpreted for schools and systems in ways that avoid a descent into 'league tables'. Whereas the project to be described in this paper has its basis in education, the techniques used will be of interest to those working with data analyses at the policy and/or professional levels more generally in the Social Sciences.

Introduction

Multilevel modeling provides a state-of-the-art method of interpreting the achievement data in students' examination results. It is the translation of the results of such modeling to a form that is straightforwardly useable by the general user that is the challenge.

Multilevel modeling copes well with sets of data which are inherently nested, where the within-group patterns of variation are of as much interest as the between-group. Students undertake study within classes within schools within courses. The between-group pattern in the analysis of examination data can show differences between schools. The within-group pattern can show whether the school is relatively more effective with higher- or lower- achieving students.

In interpreting student results, schools are most often interested in inferring educational effectiveness. Any set of achievement data is strongly influenced by factors outside the school's sphere of influence, such as gender, prior achievement and socio-economic status. Where schools have only comparison with state average as a basis for interpreting data, these factors can mask any information about educational effectiveness. Enrol high-achieving girls from wealthy families, and the school's results will be impressive. The question remains, are they as impressive as they could be reasonably expected to be?

The reporting of educational data rarely takes account of uncertainty in the measures. Without a quality analysis of data, tiny differences in comparison with previous years, other schools, other courses or the state average can be celebrated or decried, when they are no more that statistical noise. A proper comparison analysis of data should be presented with some measure of standard error in order to ensure that the conclusions draw are valid.

In NSW, students in Year 12 at the end of their high school course undertake the Higher School Certificate usually in five, six or seven courses. There is no single course which is compulsory or common to all students, although each student is required to take one of the English courses. All students in Year 10 undertake a common set of School Certificate examinations in English-Literacy, Mathematics, Science and Australian History, Geography, Civics and Citizenship. These provide a common measure of prior achievement for each of the HSC courses two years later. Obviously, the variance analysis indicates differing relationships between each of the HSC courses and the earlier SC examinations. Nevertheless,

multilevel models can account for a large portion of the variance in HSC course results, and of the variance in the aggregate of students' results over all courses.

There are 126 Catholic secondary schools in NSW who present students for the HSC. Involvement in the Data Analysis project described in this paper is voluntary, but since 2001 (the second year of the project) all 126 schools have chosen to be involved, contributing the data from their schools for analysis and receiving a report on the individual courses and students within the school. The project is conducted under the auspices of the Catholic Education Commission (NSW) as the coordinating body, and with the cooperation of the NSW Board of Studies, whose assistance is acknowledged and appreciated. Each year, following release of individual school analyses, there is a series of seminars which support interpretation of the analysis.

The set of issues described in this paper are not unique to education. Health, transport, business, law enforcement and many other areas of the social sciences require valid and in-depth statistical analyses of data to be presented in clear and useful ways to end-users who have little familiarity with the techniques being used.

Methodology

A typical multilevel model produced using MLwiN (Goldstein, Rasbash, Plewis, Draper, Browne, Yang, Woodhouse & Healy, 1998) can be depicted in the form:

$$\text{zschmkk}_{ij} \sim \text{N}(XB, \Omega)$$

$$\text{zschmkk}_{ij} = \beta_{0ij}\text{const} + \beta_{1j}\text{zscen}_{ij} + \beta_{2j}\text{zscma}_{ij} + \beta_{3j}\text{zscsc}_{ij} + \beta_{4j}\text{zschi}_{ij} + \beta_{5j}\text{zscge}_{ij} + 0.069(0.008)\text{gender}_{ij} + 0.016(0.014)\text{zseif}_{j} + 0.010(0.010)\text{zSEI}_{ij}$$

$$\beta_{0ij} = 0.050(0.013) + u_{0j} + e_{0ij}$$

$$\beta_{1j} = 0.132(0.015) + u_{1j}$$

$$\beta_{2j} = 0.168(0.014) + u_{2j}$$

$$\beta_{3j} = 0.066(0.015) + u_{3j}$$

$$\beta_{4j} = 0.239(0.014) + u_{4j}$$

$$\beta_{5j} = 0.131(0.014) + u_{5j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \\ u_{5j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.015(0.003) & & & & & \\ 0.000(0.002) & 0.003(0.003) & & & & \\ -0.001(0.002) & 0.000(0.002) & 0.004(0.002) & & & \\ 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & 0.000(0.000) & & \\ -0.002(0.002) & 0.002(0.002) & 0.002(0.002) & 0.000(0.000) & 0.001(0.002) & \\ -0.001(0.002) & -0.003(0.002) & -0.002(0.002) & 0.000(0.000) & 0.000(0.002) & 0.005(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.145(0.003) \end{bmatrix}$$

*-2*loglikelihood(IGLS)* = 4025.885(4141 of 4141 cases in use)

Fig 1: Example of a resolved multilevel model

This relates the HSC result of student *i* in school *j* in course *k* (zschmkk_{ij}) to a constant for the course, the student's results in SC English-literacy, Mathematics, Science, Australian History, Australian Geography, the student gender, the student-postcode-average socio-economic index (Australian Bureau of Statistics, 2004) and the school-average socio-economic index (Farrish, 2004). The coefficient for each of the SC results and the constant is allowed to vary at the second (school) level.

Note that it is not possible to undertake the analysis of this particular dataset using multivariate analysis because of the diversity of the student course choice. Typically 40 – 50 individual course models resolve. Of these, each student undertakes only five to seven courses. With this amount of missing data, multivariate models will not resolve.

The presentation of information from an analysis such as this for use in schools is done graphically, rather than by use of the model's equations or numerically. Graphical presentation is also a useful way to emphasize the confidence limits in the data, which can be easily glossed over in a numerical presentation.

Primary analysis

The central concept in the presentation of the analysis to schools is the comparison of the "Achieved" with "Typical" results. This is referred to as the "Primary" analysis, and the comparison is referred to as Comparative Learning Gain (CLG). The Achieved result is the mean result gained in the course by students of that course within the school. The Typical result is the mean result gained by "comparable" students; i.e., students with the same SC results, gender and SES. The Typical result is derived from the model as:

$$\hat{y}_{jk} = \overline{\beta_0 + \beta_1 x_{Eij} + \beta_2 x_{Mij} + \beta_3 x_{Sij} + \beta_4 x_{Hij} + \beta_5 x_{Gij} + \beta_6 g_{ij} + \beta_7 S_j + \beta_8 s_{ij}} \qquad (1)$$

95% confidence limits are derived from the standard errors of the aggregate.

The comparison of the Achieved with the Typical result is presented as shown in Fig 2.

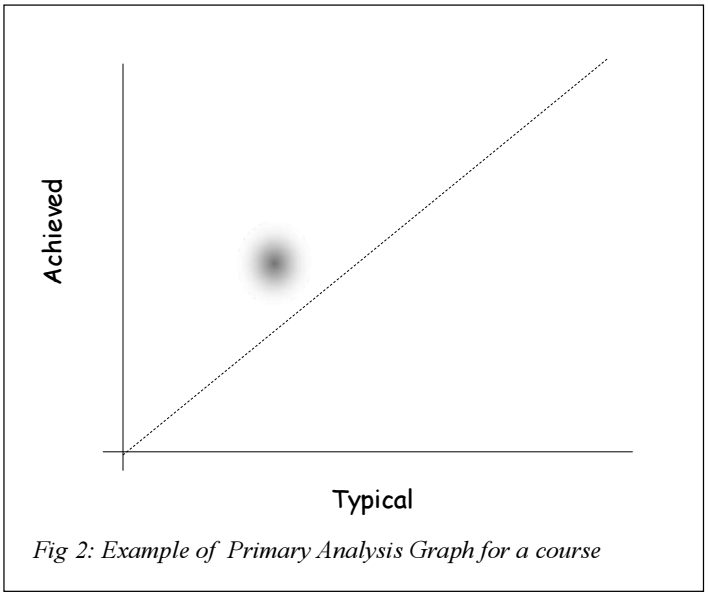


Fig 2: Example of Primary Analysis Graph for a course

The example shown indicates a result the school could be well pleased with. The result achieved in this course (indicated by the centre of the result ellipse – the "fuzzy blob") is significantly above that which would be typically gained in others of the 126 schools by students in the same course with the same SC results, gender and SES measures. The school might infer from this that either the teaching of the course or some other factor within the school is related to this positive effect.

Note in Fig 2 that the axes of the graph are not labeled. This is simply to focus the reader's attention on the relationship between the result ellipse and the Achieved = Typical line. The manual which accompanies the Analysis (DeCourcy, 2006a) indicates to the user that the axes are on a scale of +/- 2 σ. With all courses plotted on a common scale derived from the SC, we can further infer from this example that although the students undertaking the course in this school were considerable below the average of the entire cohort, they achieved a near-average result in the HSC. In an analysis which looks only at the comparison of HSC and state average, such a result might be considered inadequate. Consideration of the intake characteristics (which the teacher in the HSC course has no control over) indicates that this result is actually quite positive.

Second level effects

Second-level variation of the model contains further information of interest to the user. The first-level variation presented in the primary analysis shows simply the difference between the achieved mean of the school group and the mean which would be typical of similar students. Second-level residuals indicate the extent to which a particular school group varies from the norm in the slope of the regression line. A significantly positive second-level residual indicates that the school is relatively enhancing the results of its higher-achieving students compared to its lower-achieving. A significantly negative second-level result indicates the opposite.

Both possible results are of interest, and neither represents a pedagogically preferable outcome. While it is clearly desirable in all cases that first-level results be positive, a negative second-level result can indicate that the learning-support program for students with difficulties is being successful.

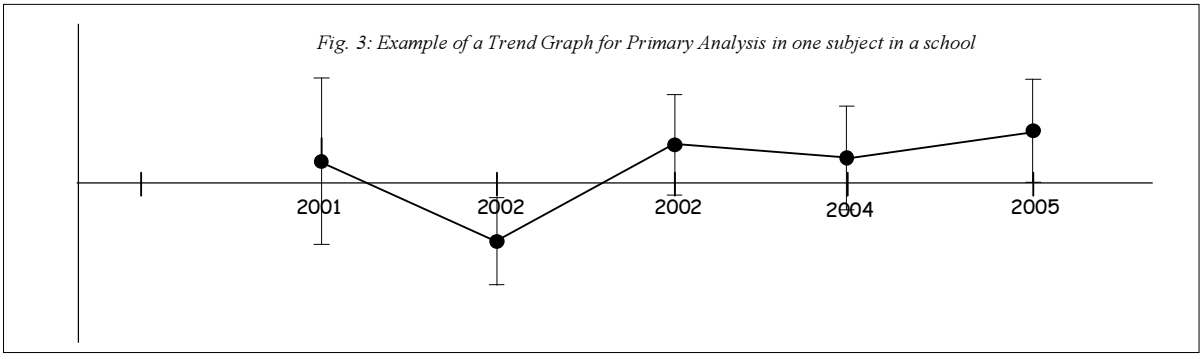
Second-level results are represented to the user of the analysis by a superimposed "+H" for a significantly positive second-level residual, and a "+L" for significantly negative. While obviously a "+H" is the same effect as a "-L", the boundary condition (first-order result greater or less than zero) for where such a change in the representation might be made is too rigid to make such a change meaningful. Hence all significant second-order results are represented simply as "+H" or "+L".

To extract the second-level residuals an "aggregate" model is constructed for each course, the graphing function of MLwiN (Goldstein et al, 1998) is used to plot a graph of the second-level residuals against school ID. This results in the construction of an exportable table of residuals and standard errors of residuals which can then be related to the primary results in each school.

Trend analysis

Once a user has considered the results in a course for a given year, the next question that arises is whether the pattern is typical of the school over time. These results are presented as a "Trends" Analysis:

Fig. 3: Example of a Trend Graph for Primary Analysis in one subject in a school



Derived from the same calculation as the primary analysis, the trend graph shows how the primary analysis in this course in this school has changed over time.

In introducing the analysis to users who are not statisticians, the trend graphs are the most powerful in providing recognition of real known factors operating within a school. The one above, for example, was of a subject taught in all five years by the same teacher whom the principal regarded as particularly effective. In 2002 she had had a period of significant illness, leading to the effect shown.

Aggregation to the whole-of-school level

Besides tracking results in individual courses, principals and others responsible for schools are interested in tracking whole-of-school performance. To do this, the techniques used in the calculation of the Universities Admission Index (Cooney, 2000) are mimicked, using the published parameters made available each year for each course (Cooney, 2006; Appendix 3). This enables a further model to be produced in the general form of that shown in Fig 1, but using the mimicked Tertiary Entrance Score as the achievement variable. Individual school results can then be plotted in a trend analysis similar to Fig 3.

Individual student results

Teachers engage with their students at the individual-student level. While there is some interest in the movement of the class mean, real engagement with the analysis for teachers comes when they see the contribution of individual students to the overall mean. To do this, a scatterplot of the HSC scores of the individual student results against the "typical" result derived from

$$\hat{y}_{ijk} = \beta_{0ij} x_0 + \beta_{1j} x_{Eij} + \beta_{2j} x_{Mij} + \beta_{3j} x_{Sij} + \beta_{4j} x_{Hij} + \beta_{5j} x_{Gij} + \beta_6 g_{ij} + \beta_7 S_j + \beta_8 s_{ij} \qquad (2)$$

is made. For the purposes of this graph the linear relationship between the standard scores and the achieved/typical HSC score is used to present students results (which teachers know) plotted against the HSC mark which typically a student of these SC results, gender and SES measures might have gained.

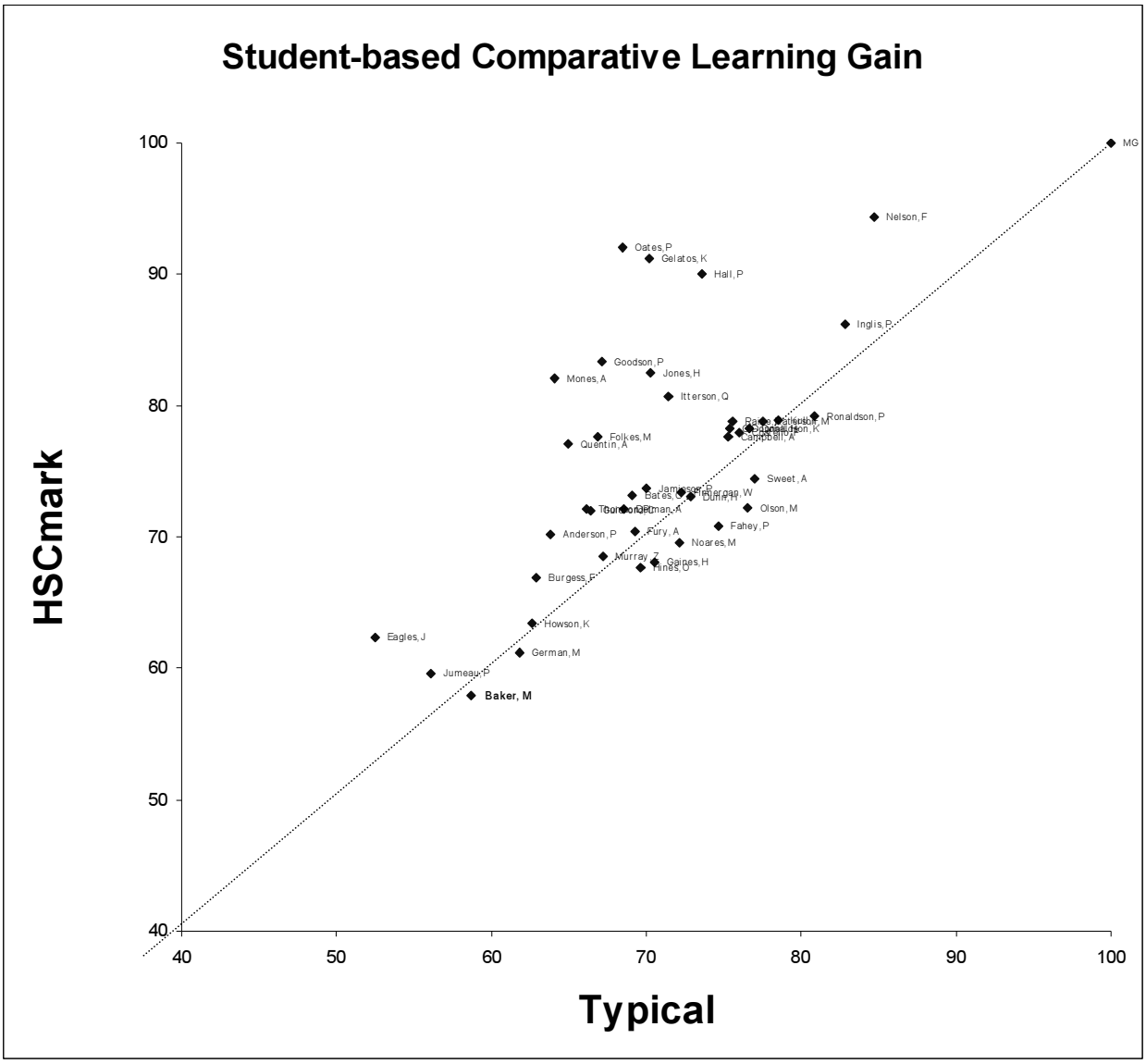


Fig 4: Example of individual-student plot of comparative learning gain.

Of course, it is not possible when plotting individual-student results to give confidence levels in the results. The point of this depiction is to enable teachers to see the elements which go to make up the overall-course results, and to construct their own diagnoses of the teaching and learning which has occurred.

Some findings from the analysis

This analysis is a voluntary exercise conducted by the agreement of systems and individual schools to pool their data for the purpose of having a sufficiently large dataset to provide meaningful multilevel analysis. An emphasis is placed on the analysis/diagnosis distinction: the package provided to each school each year is the analysis, but it is up to the curriculum experts within the school to undertake the diagnosis of the teaching/learning processes. This emphasis on the analysis/diagnosis distinction (rather than, for example, representing the analysis as a top-down accountability mechanism) has led to all 126 schools which are eligible to be part of the analysis choosing to do so from the second year of the project.

A review of the project (Catholic Education Commission, 2004) indicated good levels of user satisfaction with the service provided by the analysis. A total of 51 (41%) of users responded to the survey.

Respondents indicated use of the analysis for both informing teaching and learning

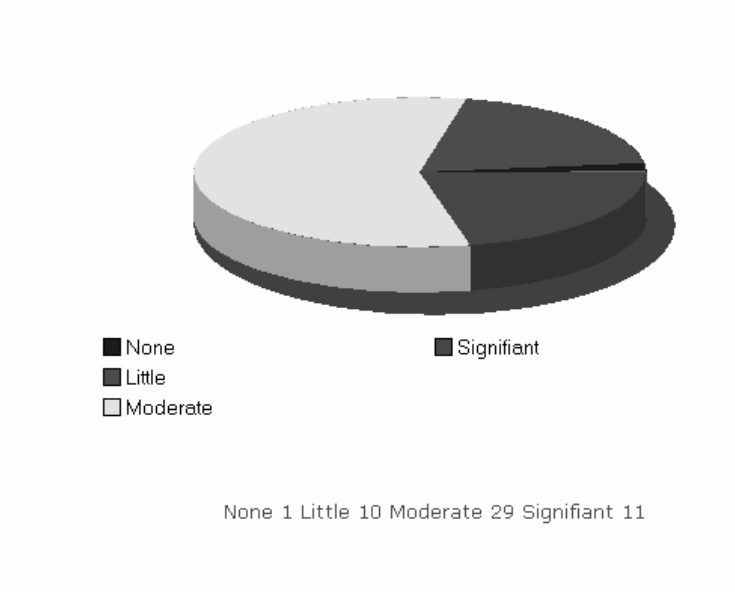


Fig 5: Respondent use of Analysis for Teaching and Learning

and also the preparation of students for the HSC:

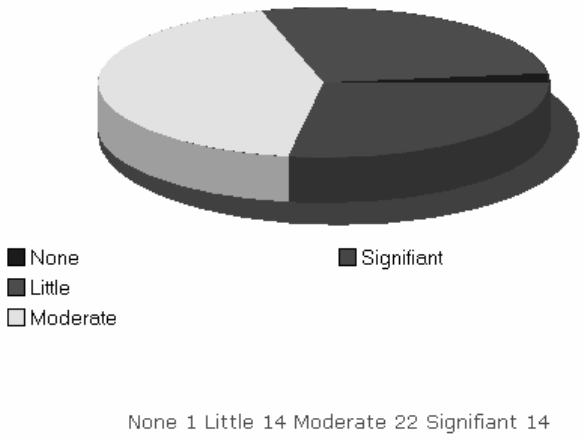


Fig 6: Respondent use of Analysis for student HSC preparation

Respondents were unanimous in wanting the analysis undertaken every year.

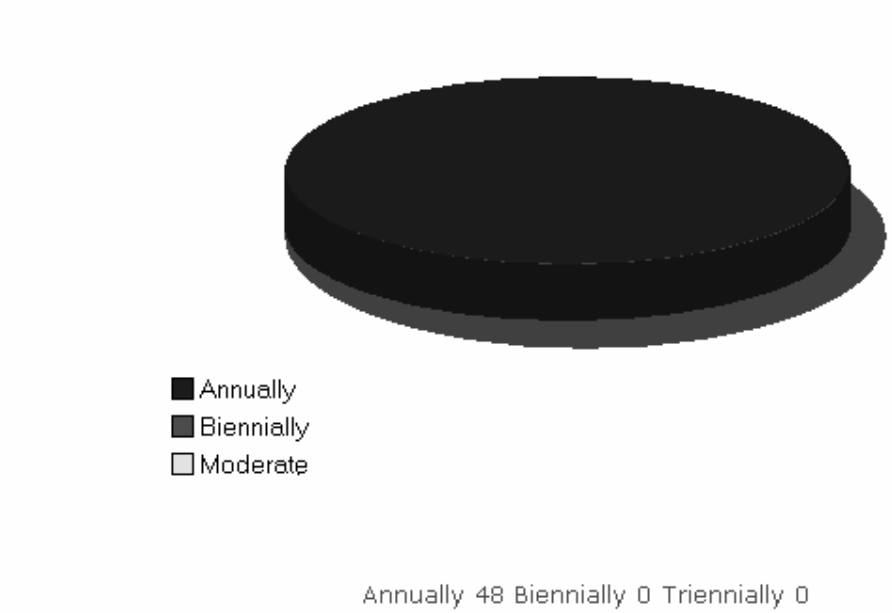


Fig 7: Respondent preferred frequency of analysis

The purpose of the analysis is its use to improve the education provided within each school. However, there are some results of interest that can be derived for a whole-of-sector or whole-of-state view.

It is clear both at the individual course level and particularly the aggregate of individual students' results that prior performance as measured by the SC test results is the factor which accounts for most of the variance in HSC results. Construction of a model where each of the variables is converted to normal equivalent deviates allows apportioning of the sources of variance, which for the 2005 aggregate Tertiary Entrance Score (TES) gave a distribution as shown in Fig 8.

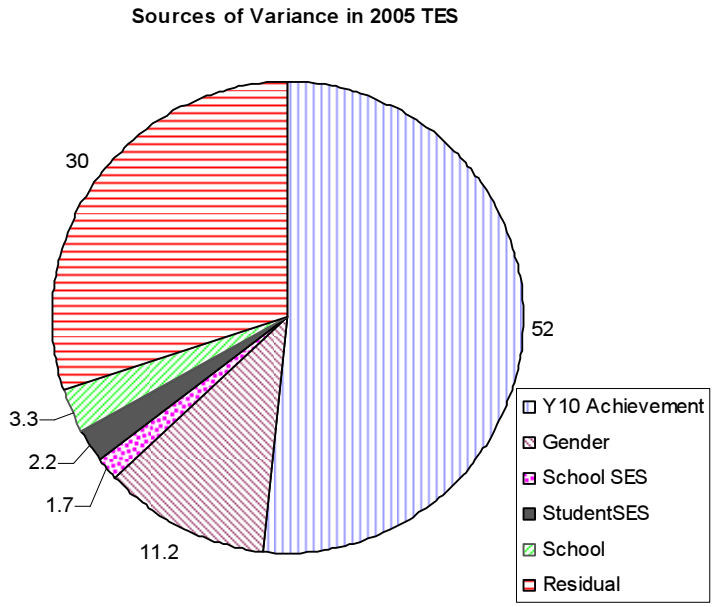


Fig 8: Sources of variance in Tertiary Entrance Score

There are a number of remarkable features of Fig 8. Firstly, more that half the variance in HSC results is related to prior achievement in the SC. Secondly, between-school variance accounts for only 3.3% of the overall HSC variance when other factors such as prior achievement, gender and SES are taken into account. Thirdly, even the use of two measures of SES accounts for a total of only 3.9% of the variance. Clearly, SES and school are much smaller factors in students' HSC achievement than is commonly thought.

This aggregate result has to be contrasted with the result in individual courses, where the between-school proportion of the variance is much larger. For example, consider the 2005 Drama results, as shown in Fig 9:

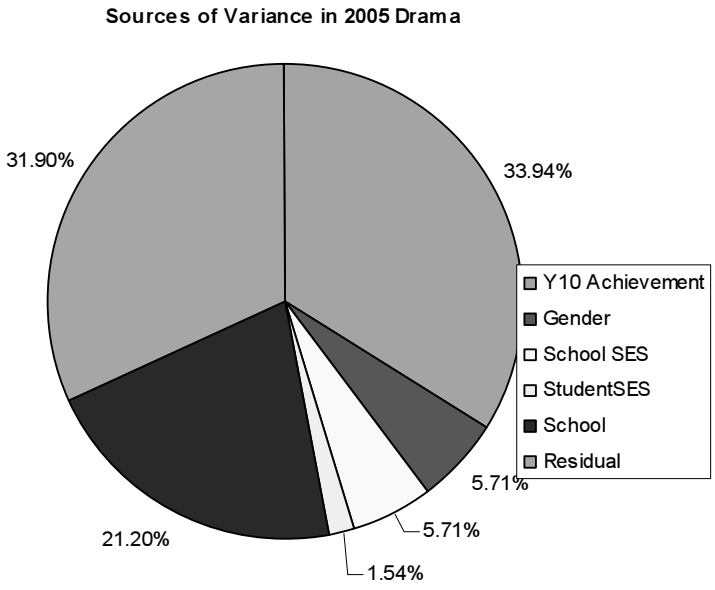


Fig 9: Sources of variance in 2005 HSC Drama

Here, between-school variance accounts for over 20%, and SES for over 7%. It is not unreasonable that we should find a much larger "school" effect in individual courses than in the aggregate of all courses for an individual student. While there is a considerable variation in teacher effectiveness, any one student is likely to have a number of teachers, some of whom are more effective than others thus canceling out each other's effect in the aggregate. The considerably larger SES effect seen in Drama than in most other subjects is harder to explain, but might be hypothesized to be based on more exposure to live theatre among higher-SES students.

The aggregating effect of gender on students' results is the opposite of school/teacher. Whereas noticeable differences in individual teacher effects aggregate to cancel each other out, the small gender effects being the same for each course of any one student, tend to accumulate. It is of interest to note that over the five years the analysis has been conducted, the gender effect initially rose, but then declined a little as shown in Fig 10.

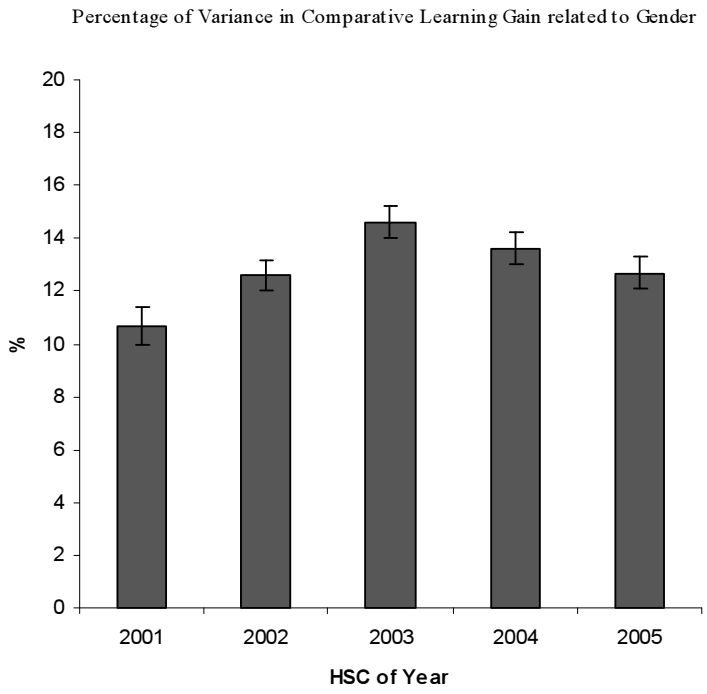


Fig 10: Gender effects in aggregate (TES) results

However, if we are to say that boys education strategies are really starting to take hold, this graph has a fair way to go before it represents a gender-neutral result.

Conclusion

Is it possible to use complex statistics in a way in which the general user can gain reasonable benefit? The anecdotal responses, the survey of users and feedback from the analysis seminars suggests that it is so. In the absence of access to whole-of-state data for all students, it is not possible to construct a proper multilevel model to test the results in the Catholic sector against all. Further, if all of the state's students are improving in their learning outcomes against standards, a real improvement might be masked by the backdrop.

However, consideration of just the last three year's worth of data (since the seminar program on the analysis began in 2002) indicates an interesting pattern in the comparison between numbers of subjects above and below mean for the sector as a whole, as shown in Fig 11.

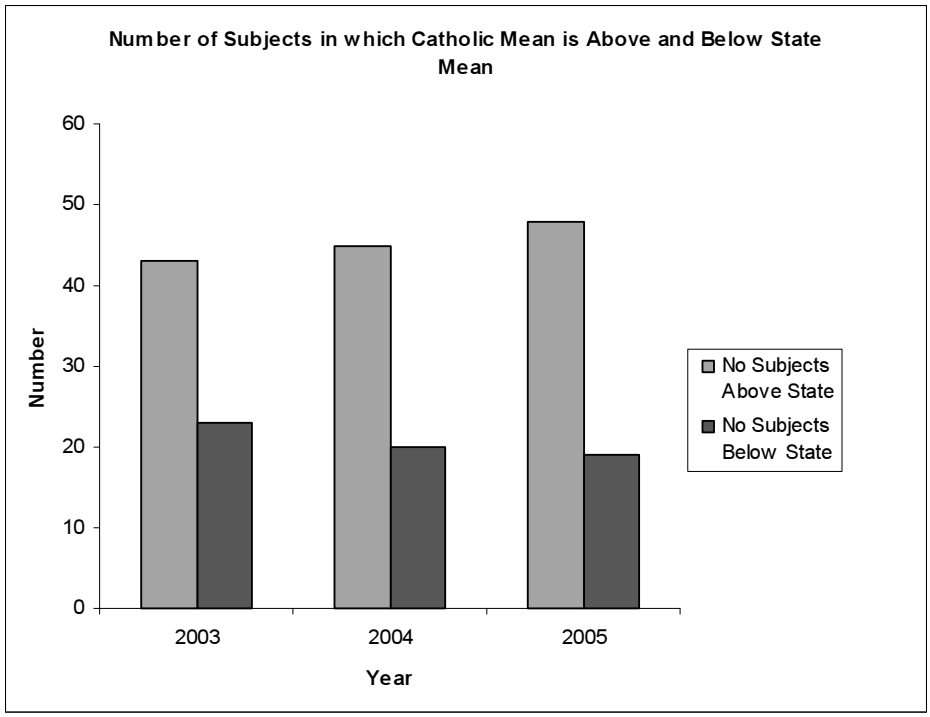


Fig 11: Catholic sector comparison with whole-of-state since beginning of analysis seminars.

An interesting facet of the 2006 analysis will be to see if this pattern is continued.

Bibliography

Amrein, A. L., & Berliner, D. C. (2003). 'The effects of high-stakes testing on student motivation and learning', *Educational Leadership, 60*, 5, 32-38

Australian Bureau of Statistics (2004). SEIFA 2001: *The socioeconomic index for Areas*. Canberra: Self

Braun, H. I., & Mislevy, R. (2005). 'Intuitive Test Theory', *Phi Delta Kappan, 86, 7*, 489-497

Catholic Education Commission, NSW, (2004) *HSC Data Analysis Project: School Survey* Sydney: Self

Cooney, G. (2000). The Universities Admission Index (UAI). Sydney: Universities Admission Centre

Cooney, G., (2006). Report on the scaling of the 2005 NSW Higher School Certificate. Sydney: Universities Admission Centre; NSW Vice-Chancellors Conference Technical Committee on Scaling

DeCourcy, J. S. (2006a). *HSC data analysis: The manual*, Sydney: Catholic Education Commission.

DeCourcy, J. S. (2006b). *Report on HSC 2005*. Sydney: Catholic Education Commission.

Farish, S. (2004). Funding Arrangements for Non-government Schools 2005-2008: Recalculation of the Modified A Socioeconomic Status (SES) Indicator using 2001 Australian Bureau of Statistics Census Data Canberra: Department of Education, Science and Training.

Goldstein, H., (1995). Multilevel Statistical Models London: Arnold, 2nd edition.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G. & Healy, M., (1998). *A User's Guide to MlwiN*, University of London: Multilevel Models Project, Version 1.0

Marks, G. N., Rowe, K.J., & Beavis, A. (2003). Australian Schools not so 'Undemocratic', *Campus Review, June/July 34*.

Marsh, H. W., (1991). Failure of high-ability high schools to deliver academic benefits commensurate with their students' ability levels, *American Educational Research Journal, 28, (2)*, 445-464.

Marsh, H. W. & Rowe, K. J., (1996). The negative effects of school average ability on academic self-concept: An application of multilevel modelling, *Australian Journal of Education, 40*, 65-87.

Marsh, H. W., Chessor, D., Craven, R. & Roche, L. (1995). The effects of gifted and talented programs on academic self-concept: the big fish strikes again, *American Educational Research Journal, 32, (2)*, 285-319.

Marsh, H. W., Hau, K. T., & Craven, R. (2003). The big-fish-little-pond effect stands up to scrutiny, *American Psychologist, 59, 4*, 269-271.

O'Day, J. A. (2002). Complexity, accountability and school improvement, *Harvard Educational* Review, *72*, 3, 293-329.

Rowe, K. J. (2000). Assessment, league tables and school effectiveness: Consider the issues and 'Let's Get Real', *Journal of Educational Enquiry, 1, 1*, 73-98.

Rowe, K. J., (2001). Equal and different? Yes, but what really matters?, Riverview Boys Education Conference, October 2001.

Rowe, K. J. (2004a). 'The Importance of Teaching: Ensuring better schooling by building teacher capacities that maximize the quality of teaching and learning provision – implications of findings from the international and Australian evidence-based research'; Paper presented at the 'Making Schools Better' Conference, Melbourne University, 26-27 August 2004.

Rowe, K. J. (2004b). '*Analysing and Reporting Performance Indicator Data: 'Caress' the data and user beware!*' Paper presented at the 2004 Public Sector Performance and Reporting Conference, Sydney 19-22 April 2004.

Smith, M. (1999). Methods for within-school comparison of subject results Pers. Comm.

Visscher, A. J. & Coe, R. (Eds), (2002). School improvement through performance feedback, Lisse, The Netherlands: Swets and Zeitlinger.