Models for reality: new approaches to the understanding of educational processes.

A professorial lecture given at the Institute of Education July 1, 1998

Abstract

Much of the dissatisfaction with existing quantitative explanations in education and the social sciences, arises from an over-simplification of real life. This talk will argue that in recent years new quantitative methodologies have been developed that provide powerful tools for studying social structures and processes. These 'multilevel' models, which attempt to describe the complexity of the real world, have begun to yield important research insights and to provide a rational basis for the critique of certain contemporary educational policies. The talk will describe the essential features and potentialities of these procedures in a non-technical fashion.

The truth about beauty

General relativity: a theory too beautiful not to be true - Albert Einstein

While Einstein seems to have been right, so far, about the truth of general relativity, his justification for it seems less secure. The applicability of general relativity is to be found in empirical evidence, not in mathematical aesthetics. Our emotional feelings about our *models* of reality may colour our beliefs and the way we act upon those beliefs, and that has often been a useful guide, at least in the natural sciences. Yet such beliefs need to be judged by a careful study of the real world. In this talk I do not wish to argue against elegance, but I do wish to question the notion that elegance, or at least simplicity, equates to truth.

In the social sciences, and especially in Education, descriptions which have possessed elegant simplicity have often also been wrong. There is a range of activity that goes from the early exponents of intelligence measurements, through the vast industry associated with most of modern psychometric modelling, through to the simplistic world of educational league tables. I will argue that in order to describe the complex reality that constitutes

educational systems we require modelling tools that involve a comparable level of complexity. I also wish to argue that, while we need continually to elaborate our models, we will almost certainly remain a long way from perfect descriptions; the journey is important, even though we may never arrive at our destination. It follows that we should also strive to provide some way of knowing how far we may be from a complete description; in other words we require a measure of our ignorance as well as a description of our knowledge.

I am, of course, talking about quantitative models. I will have little to say about non-quantitative explanations and models, yet it does seem to me that one of the reasons for the unfortunate gulf between the exponents of quantitative and non-quantitative educational understandings is that exponents of the latter tend to view the former as simplistic and reductionist. As will become clear, I have a certain sympathy with that view because I believe it has some justification in reality. One of my main purposes, however, is to demonstrate that quantitative models do not *need* to oversimplify reality in the way that they often do, and I want to suggest that they can begin to provide usefully detailed descriptions of the world, and thus perhaps prepare the ground for a reconciliation of research methodologies.

To begin, I shall illustrate my general theme by looking at the way particular models of mental testing have come to dominate certain areas of educational assessment. I will then illustrate some consequences of this by looking at a recent international comparative study of adult literacy. Following that I will describe the work which has occupied most of my own research time for the last 15 years and attempt to show how this work is leading many people to think about education, and indeed many other areas of social and biological science, in new and powerfully constructive ways that begin to capture some of the complexity to which I have referred.

Item Response models

Item response models are too good not to be true - A leading psychometrician

Unlike general relativity the evidence for this assertion is decidedly lacking. Psychometrics has enjoyed a highly privileged status within education as a mathematically based discipline which seeks to provide a formal structure for making statements about mental abilities and student achievements. To be sure, it has achieved a certain level of technical

sophistication, but on close examination this sophistication resides in the dexterity required to do the computing necessary to obtain decent numerical results from the models used to describe the data. The models themselves, as in the above quotation, have remained at a surprisingly simple level of description; so much so that they stand little chance of adequately representing the complex reality of the real world. Unlike many parts of the physical world, the social world does not lend itself, in my view, to description in terms of simple formulae. Nevertheless, what Stephen Gould once referred to as 'Physics envy' (Gould, 1981) does seem to motivate many psychometricians, and hence the above quotation.

To illustrate what I mean I shall explore a recent important survey of adult 'literacy' the International Adult Literacy Survey (IALS) (Murray et al., 1998) supported by OECD and carried out in 9 countries and involving an interview with about 3,000 adults in each. Like most major international comparative studies this one was dominated intellectually by psychometric practice from the United States. In the case of IALS its design was actually based upon three major U.S. literacy surveys, which influenced the aims and content. From the outset it was decided that there were three functional literacy proficiency 'domains': Prose literacy, Document literacy and Quantitative literacy (numeracy) (Murray et al, 1998). For each participant in the study, a proficiency 'score' for each of these domains was estimated from responses to a set of tests or tasks. These scores then formed the basis of international comparisons.

A considerable controversy arose towards the end of the study with one country (France) withdrawing completely after it emerged that it had the lowest scores on all three domains, followed by the EU setting up a project to re-evaluate the results. As with all international comparisons of competence or achievement a fundamental issue is whether it makes much sense to use a single common set of tasks in a variety of very different cultural, educational and social settings. IALS itself addresses this issue in a very limited study comparing the 'difficulties' of some tasks for their French and English translated versions. One conclusion is that the necessities of translation make tasks more or less difficult in different contexts. Similar findings about the incommensurability of translated materials have been obtained by others (see Goldstein, 1995 for a summary), and the use of common measuring instruments therefore raises the issue of who is advantaged and who is disadvantaged in the process. I shall not go into this issue here, except to remark that it is perhaps the most important one yet to be addressed in the field of international comparisons.

My concern, rather, is with the way in which the crude psychometric steamroller squeezes such considerations to the periphery of technical appendices which, I suspect, few will ever read.

For each proficiency in IALS there is an 'item response model', a modern day variant of factor analysis, which makes the really simple assumption that the responses to the constituent tasks are all determined by just one underlying 'factor' or ability or proficiency - call it what you will. Interpreting this literally some 10% of the tasks were excluded on the grounds that they didn't fit such a model. A proficiency score for each individual was then calculated using a weighted average of the responses (correct/incorrect) to each remaining task.

This produces just three numbers for each individual describing their literacy proficiency, with the assumption that these are directly comparable across countries. The psychometric model has formed a procrustean framework that excludes or downweights some components that don't fit its simplistic assumptions. Nor is this an isolated example; this kind of psychometric reductionism is hugely popular among those devising computerised testing procedures where the use of such models greatly simplifies the whole exercise and creates an appearance of precision and objectivity.

Another interesting example of the attraction of such simplistic approaches occurred in the late 1970s when the government's Assessment of Performance Unit (APU) was concerned to determine whether average standards of achievement were changing over time. For those who are interested the episode has been fully documented by Caroline Gipps and myself in a study of the APU carried out in the early 1980s (Gipps and Goldstein, 1983).

The NFER at that time was promoting the so called 'Rasch model', the simplest of the item response models, as the technique that could provide an answer to this question. After much debate, and also a great deal of technical obfuscation, the APU, albeit reluctantly, accepted that there was no acceptable way of measuring absolute trends over time. To summarise, it turns out that there is no objective way to separate 'real' changes in student performance from changes in test difficulty. This earlier debate bears a striking similarity to current government proposals, which advocate the achievement of specific targets to be reached at Key Stage 2 over the next few years; one can only hope that those politicians willing to stake their careers on such notions are familiar with recent history.

At this point I suppose the obvious question to ask is 'why have such simple-minded models persisted when more complexity can be introduced'? A complete answer to that would, I think, make an interesting historical study, but let me make some interim suggestions which will also lead me on to the main concern of this talk.

The most obvious feature of the statistical models I have been describing is their apparent technical complexity, despite their simplistic approach to their subject matter. This immediately puts them beyond critical comment for the vast majority of subject matter specialists, in literacy or whatever, who have no means of understanding the technicalities. This sets out a fertile ground for the psychometrician to dominate the debate, invoking the high status generally associated with mathematical reasoning. Powerful commercial interests in the shape of the, largely U.S., testing agencies are also important here since it is very much in their interests to act as providers of sophisticated know-how. As I have attempted to illustrate in the IALS case, however, the sophistication lies in the details of the calculations associated with obtaining proficiency scores and the like, and *not* in the complexity with which the real world is described. Once the emphasis shifts to attempting to capture real world complexity in the setting up of mathematical models, then the subject matter specialists have to be let in on much more of the act and the complexity of the computations tends to become less important. To be sure, subject matter specialists are involved in designing the questions and tasks, but thereafter they assume a much more passive role. If they are brave enough to suggest that some complexities have been overlooked then they may well be dismissed as having not properly understood the technicalities (for an example see Street (1996) and Jones (1996) in the context of IALS).

The other point, of course, is that it is really quite difficult to provide mathematical or statistical models which do begin to approach the complexity of the real world, and even where this can be done, the costs of obtaining adequate data to test out these models is very high. This is, however, no excuse for perpetrating inadequate models as if they were realistic descriptions. My own view is that such promotion has been an important cause of the polarisation within education, and the social sciences more generally, between the quantitative and qualitative schools of research. Yet, to argue against the over-selling of certain simplistic views of the world, is not to argue against the application of quantitative methods. On the contrary, the use of quantitative methodologies which are rich enough to match the real complexities of the social world may

eventually allow us to bridge the gap between qualitative understandings which emphasise those complexities, and quantitative tools and understandings which provide formal descriptions of them which can be operated upon to obtain testable predictions, further refinements and useful generalisations.

I am going to describe some of the work that myself and colleagues have been carrying out in one area where we have been developing statistical models of increasing complexity to describe educational, social and other systems. This is not the only area where complex models are likely to prove fruitful, it just happens to be one where I think many people have now been persuaded that these models are the appropriate ones to apply. I refer of course to the area of multilevel modelling with which the Multilevel Models Project at the Institute of Education has been engaged for some 12 years. The remainder of this talk deals with these models and I must make it clear that the credit for any achievements that have been made must be shared among the members of the project team as well as numerous researchers who have contributed to this field around the world. Nor in this talk can I possibly cover all of the work that is currently going on, including the particularly exciting advances in user software interface design being developed by my colleague Jon Rasbash.

Complex models for complex systems

Social systems, and in particular educational structures, are built around identifiable groups of individuals, whether these are families, schools, neighbourhoods or friendship groups. These groupings are purposeful; they share opinions, attitudes or achievements. In other words the individuals within them have characteristics which derive from the natures of the groups they belong to as well as from their own individual endowments and histories.

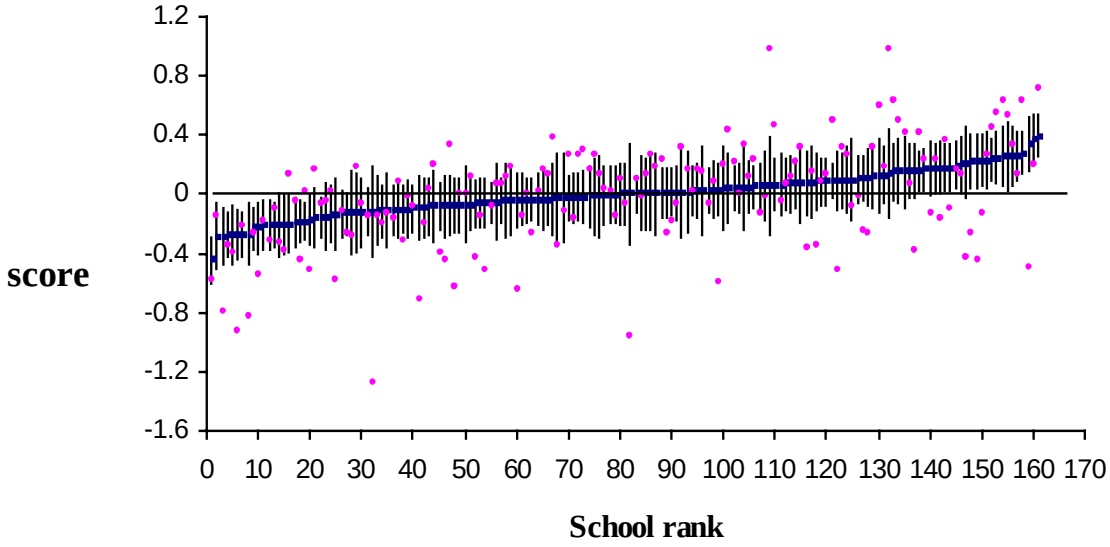
In schools children learn and achieve as a result of what they themselves bring but also as a result of what the school provides, through its teaching, physical facilities and ambience. Children also learn from their peers and the characteristics and achievements of the other pupils in the school will impinge on the development of any particular child. If we wish to capture such complex relationships then we need to have tools which are capable of matching that complexity. What I hope to do is to persuade you, through an example, that it is possible to do this, to thus obtain new insights into

schooling, and to convey the results in a way that can be appreciated without a detailed technical understanding.

We and others first began to develop the statistical models and the computer software that allowed us to explore hierarchical data structures in the mid 1980's. In education these hierarchical structures were initially conceived as consisting of pupils grouped or nested within schools, themselves grouped within Local Education Authorities. Schools differ in terms of the average achievements of their students, partly because those students have different achievements when they enter the schools - because of where the school is situated or because they operate a selective intake policy as well as because of the learning experiences which differ from school to school. The first major multilevel analysis of educational data (Aitkin and Longford, 1986) was also a prototype for subsequent developments in school effectiveness modelling, and the principal developments in that area would not have been possible without the ability to fit these models.

In brief, what these models allowed researchers to do was to quantify how much of the variation in achievement among pupils at any stage of schooling could be attributed to the school, after adjusting for pupils' initial achievements - the so called value added model of schooling. The finding that anything up to 15% appeared to be due to schooling encouraged more detailed studies trying to disaggregate this figure for different groups of pupils such as those from different ethnic and social groups and for those with different levels of initial achievement. Work with large samples (see for example O'Donoghue et al, 1997) has shown that schools differ considerably, for example with some appearing to perform relatively well for low achievers but not for high achievers. Up to a point these models also allow us to identify the contribution from individual schools - their so called 'value added scores'. Yet we now know from a number of studies that estimates of these contributions have so much statistical uncertainty attaching to them that it is impossible reliably to make valid comparisons between most schools. It is this finding above all that provides strong evidence against current policy on the publication of league tables - of whatever kind, and those who advocate evidenced based policy making would do well to understand this. The following graph (Yang et al., 1998) illustrates this where, roughly speaking, schools can only be separated statistically if their intervals do not overlap.

KS1 reading 'value added' uncertainty intervals + unadjusted means (dots)



A kind of uncertainty principle operates; we can describe structural relationships at a high level of complexity but we cannot precisely measure the contribution from each individual school.

As familiarity with this kind of modelling has grown so has the realisation of its limitations. Real life data are rarely structured in purely hierarchical ways. Pupils are grouped within schools, but they also are grouped by the neighbourhoods where they live and which may themselves exert some influence on their learning. If we are studying secondary schooling then public examination results at KS4 are influenced not merely by the secondary school but also by the Junior school attended by the pupil (Goldstein and Sammons, 1997). Further, many pupils move school so that the effects of schooling will involve contributions from all the schools attended. Another way of thinking about such situations is that many pupils are members of more than one school and models which allow only simple hierarchical structures are not complex enough to describe this.

Schools themselves are complex organisations and increasingly attention is turning away from the school to the level of the classroom with all the complexity involved in describing how such influences combine throughout a pupil's schooling career (Hill et al., 1995). What we have

found is that the original models that we started with can be extended to handle such complexities. Thus, they can deal with several outcomes simultaneously, for example with achievements in different subject areas or mixtures of achievements, attitudes and behaviour measurements. Different kinds of outcomes can be handled, such as examination grades, key stage levels as well as continuously distributed test scores and mixtures of these simultaneously. The contributions of the several schools attended by a pupil can be modelled so that, at least in principle, pupil mobility can be taken into account. Work is currently going on into ways of making adjustments for errors of measurement and missing data and the main limitations appear to be the practical ones associated with the availability of computing power, and, crucially, high quality and extensive data, rather than a lack of methodological tools. I should also mention that these complex models are also being applied in many other areas, such as the analysis of the spatial distribution of disease, the provision of health care services, the analysis of fertility patterns and the study of salmonella in chickens to name just a few.

We are also engaged in research, led here by Min Yang and in collaboration with Simon Thompson and colleagues at Imperial College, into 'meta analysis', a term used for procedures which combine data from separate research studies in order to produce a more reliable conclusion. These procedures can all be regarded as special kinds of multilevel models and when this is done, not only can efficient combination procedures be devised, but new possibilities emerge. Thus, for example, some studies may report simply average differences between categories or other summary statistics such as regression coefficients, whereas for others individual records may be available for reanalysis. Such data can be combined efficiently within a single model, and this greatly extends the possibilities for this kind of analysis.

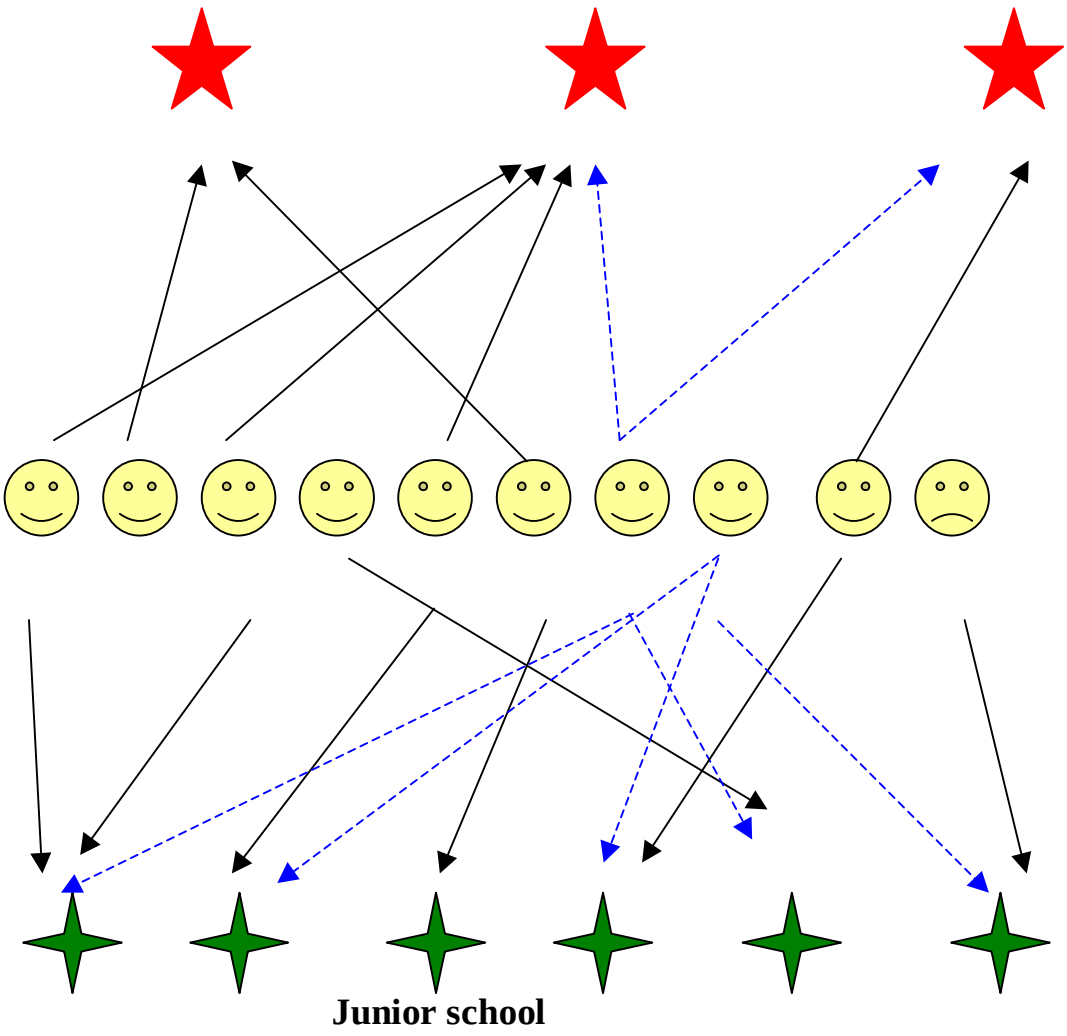
To illustrate some of what I mean by all this I shall quote from recent work carried out together with Pam Sammons on the effects of secondary and of junior school attended on GCSE examination results.

Junior and Secondary school influences on examination results

The ILEA Junior School Project (Mortimore et al., 1988), which remains one of the most important school effectiveness studies, followed up children in 50 Junior schools and subsequently these were followed through secondary school with information on their GCSE results. The

pupils had measures of achievement in Reading and Mathematics at the beginning of Junior school, on reading and (grouped) verbal reasoning at the end of Junior school and a GCSE total point score based on a simple scoring system which assigns a score of 7 to an A grade, 6 to a B etc.

A conventional 'value-added' analysis of GCSE scores with adjustments for the 11 year old (end of Junior) reading score and verbal reasoning group, shows that about 6% of the variation in GCSE scores can be attributed to the secondary school attended. When the pupils are also identified by their junior school, and achievement at the start of junior school is adjusted for, this drops to 2% whereas that associated with the junior school attended is 6%. In other words the junior schools appear to account for three times as much of the GCSE score variation than the secondary schools. Now to some extent this finding may not be as surprising as it first seems. Secondary schools are some 3 times as large as junior schools and for this reason alone one would expect the variation between them to be less than that for smaller institutions. Nevertheless, whatever the reason, it seems clear that an analysis which ignores junior school membership provides an incomplete description of secondary school effects, and we should also remember that the GCSE score is just a total score and not disaggregated into subjects. So far few people have attempted to replicate these results but recent work by Sally Thomas and her colleagues suggests that there are situations where our findings do not hold and that Secondary schools can be associated with *more* variation in GCSE scores than Primary schools.



Multiple membership

Single membership

The Figure illustrates the cross classification, and also the 'multiple membership of junior and secondary schools, and in one case a pupil whose

secondary school identification is unknown. It turns out that, if we are prepared to assign probabilities of belonging to each secondary school for such a pupil then we can carry out a valid analysis. This latter feature allows us to handle data where , for example, group identifications may be lost or simply unavailable, but where the known possibilities can be assigned membership probabilities. For a further discussion see Hill and Goldstein (1998).

Thinking about complex structures

The existence of modelling tools which can deal with complex structures has two further implications. First it should encourage researchers to collect data in such a way that this complexity can be studied. If we wish to look at the influence of peer group characteristics on individual achievement we need to collect detailed longitudinal data on groups of pupils learning and developing together. These groups will not be neatly defined in terms of classes, but will form and reform within and outside the school; as social groups of all kinds. It is just this level of complexity that we now have the techniques to deal with.

Secondly, the availability of the technical tools gives us new ways to think about social data. Traditional statistical procedures such as regression analysis, for the most part treat individuals as independently acting entities. Membership of interlocking group structures is ignored and possible causal effects are studied for their influence on an individual. As I have argued, this picture of reality is at best incomplete, and at worst leads to distorted conclusions. If we start from the assumption that social structures and group memberships are influential then we not only will begin to design studies in different ways, we will also begin to think about causality differently. We will begin to see individual actions as mediated by those of others in whom they are in contact and by the institutions of which they are members.

Finally I want to talk about the insights which this thinking about complex structures has given us into a very old debate in the social sciences, namely the role of randomised experiments versus the collection of data as they actually exist without randomly assigning individuals to different 'treatments', for example teaching groups with different reading schemes.

It has long been assumed that the gold standard with respect to obtaining sure knowledge about causality lies with randomised trials. In medicine this is expressed in strict requirements for carrying out clinical trials for new

drugs etc., and there is a body of opinion, within those who strongly advocate evidenced based medicine, that this is the only really sure way to knowledge. In a recent paper (Goldstein and Blatchford, 1998) Peter Blatchford and I argue that this is not necessarily the case, and that there are theoretical as well as practical reasons for viewing randomised controlled trials as inferior in some circumstances to observationally based studies. The example I shall use is that of studies of the effects of class size on progress in achievement, but it will apply in a wide variety of circumstances.

Consider setting up an experiment where, at random, some children are assigned small classes and others large classes, the children followed up over a sufficiently long period at the end of which they are tested to see which group made the most progress. This is the classical randomised controlled trial. The standard argument for randomisation is that if we can detect a difference in achievement in favour of the small classes, then we can be sure that, at least on average, we are justified in concluding that small classes produce greater progress in reading. In particular cases, however, where 'compositional' effects operate this conclusion does not follow.

For the sake of argument suppose that class size itself has no effect on progress. Suppose also that there is an 'effect' due to the proportion of a *particular group* in the classroom, say low attaining children, and consider the extreme case where the *only* effect on progress is where the percentage of low achievers is at least 33%, and that only when this is exceeded does the progress of the students in that class become lowered. Assume that the average proportion of low achievers in the population is 10% and that the proportion of small classes with this percentage of low achievers is the same as the proportion of large classes. Assume also that the low achievers are clustered in classes in such a way that classes tend to contain either a large proportion of low achievers or very few low achievers. In a reasonably large random sample of classes, therefore, we would expect to obtain useful numbers of small and large classes with high proportions of low achievers.

If we carry out an RCT and given that just 10% are low achievers, then for a small class of size 15 the probability of obtaining a class with at least 33% low achievers is rather small, and, importantly, decreases dramatically as the size of class increases. This is illustrated below.

Low achiever compositional effects

Percentage of low achievers in population = 10%

With random allocation:

Class size	Prob. of class $\geq$ 33% low achievers
15	0.013
20	0.0023
25	0.0005

In a study with 200 classes size 15, and 200 classes size 25:

Expected percentage no 'low achieving' class among large classes = 90%

Expected percentage no 'low achieving' class among small classes = 7%

While, in a reasonably large study with 200 small and 200 large classes, it would be common to find one or more small classes with a high proportion of low achievers it would be very uncommon to find any large classes at all with a high proportion of low achievers. In such a study a comparison between large and small classes might detect an effect simply because the small classes were the only ones where the percentage of low achievers was at least 33%. Thus an 'effect' of class size would be simply an artefact of the random sampling procedure, and any causal inference would be incorrect.

We see here an instance of where a key rationale for randomisation, namely the equalisation (on average) of initial characteristics within the 'treatments' being studied, undermines the possibility of valid inferences. On the other

hand, given the assumptions I have made, an observational study where existing large and small classes were sampled, would allow us to avoid coming to an erroneous conclusion. The principle of randomisation in this case, by creating an artificial context, prevents us from studying causality.

This example illustrates one way in which an appreciation of the complexity of educational data can begin to inform basic notions about design and causality as well as procedures for data analysis. Although it is beyond my brief today I think it would be interesting to speculate on how far this example from education may apply in other fields, such as medicine and biology.

Conclusion

My thesis has been that to describe complex educational and social realities we need to use modelling tools that attempt to match that complexity without, of course, confusing the model with the reality itself. As my psychometric examples seek to demonstrate, mathematical complexity for its own sake is insufficient. Likewise, there is no justification in using a procedure *because it is simple*. I have attempted to show how some quite complex realities can not only be modelled, but also how the application of these models has provided ways of describing the world in terms that can be widely understood.

The following quotation from a recent DfEE consultation document on 'performance tables' (DfEE, 1998) illustrates how widespread, nevertheless, is the desire to ignore complexity.

In arguing for a particularly crude form of value added analysis of GCSE results the document finds virtue in the fact that "it does not depend on a complex mathematical model". On the contrary, what we should aim for are descriptions which are at the level of complexity which is appropriate to the system being studied, and in so doing advance our qualitative as well as quantitative understanding of education.

Let me close by saying that it has been a great privilege to work at the Institute of Education. In the face of an often hostile climate, the Institute has been able to retain an ethos of critical scholarship, recognising the complexity of the real world and maintaining a resistance to the simplicities of political fashion .

Harvey Goldstein
Institute of Education

References

Gipps, C. and Goldstein, H. (1983). *Monitoring Children*. London, Heinemann.

Gould, S. J. (1981). *The Mismeasure of Man*. New York, W. W. Norton

Murray, T. S., Kirsch, I. S. and Jenkins, L. B. (1998). *Adult literacy in OECD countries*. Washington, DC, National Center for Education Statistics.

Goldstein, H. (1995). *Interpreting international comparisons of student achievement.* Paris, UNESCO.

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

O'Donoghue, C., Thomas, S., Goldstein, H. and Knight, T. (1997). *DFEE study of value added for 16-18 year olds in England*. London, Department for Education and Employment, & Institute of Education. (available at http://www.ioe.ac.uk/hgoldstn/#download).

Goldstein, H. and Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. *School effectiveness and school improvement.* **8**: 219-230.

Jones, S. (1996). Ending the myth of the 'Literacy Myth'. *Literacy across the curriculum* **12**: 17-26.

Street, B. (1996). Literacy, Economy and Society. *Literacy across the curriculum* **12**: 8-15.

Hill, P. W., Rowe, K. J. and Holmes-Smith, P. (1995). *Factors affecting students' educational progress: multilevel modelling of educational effectiveness.* Intl. Congress for school effectiveness and improvement., Leeuwarden, Netherlands.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., et al. (1988). *School Matters.* Wells, Open Books:

Hill, P. W. and Goldstein, H. (1998). Multilevel modelling of educational data with cross classification and missing identification of units. *Journal of Educational and Behavioural statistics (to appear).*

Goldstein, H. and Blatchford, P. (1998). Class size and educational achievement: a review of methodology with particular reference to study design. British Educational Research Journal, (to appear).

DfEE (1998). Secondary school and College performance tables: A consultation document. London, DfEE (March 1998).

Yang, M., Goldstein, H, Rath, T. and Hill, N. (1998). The use of assessment data for school improvement purposes. (submitted for publication)