In *International Journal of Social Research Methodology,* Vol. 8, No.3 July, (2005) pp225-238 *ISSN 1364-5579*

# Students' Developing Knowledge in a Subject Discipline: Insights from combining Quantitative and Qualitative Methods

**Celia Hoyles[1], Dietmar Küchemann[2], Lulu Healy[3] and Min Yang[4]**

*Abstract*: In this paper we describe an example of research that combined quantitative and qualitative methods in order to investigate students' developing mathematical reasoning over time and to identify factors that were influential in this development.

[1] School of Mathematics, Science & Technology, Institute of Education, University of London
[2] School of Mathematics, Science & Technology, Institute of Education, University of London
[3] Programa de Estudos Pós-Graduados em Educação Matemática, Pontifícia Universidade Católica De São Paulo, Brazil
[4] Forensic Psychiatry Research Unit, Institute of Community Health Sciences, Queen Mary University of London

## Setting the scene: investigating the development of mathematical reasoning over time

We describe an example of research adopting an approach that combined quantitative and qualitative methods in order to investigate students' developing mathematical reasoning over time. The study therefore complements the growing corpus of studies that have employed multiple methods (see, for example, Brannen, 1992). We describe the outcomes of using the different approaches, the strengths and the challenges of each method and the benefits of mixing them at different points in the study.

A fundamental objective of mathematics education must be to help students recognise and construct mathematical arguments. Yet it is well known in mathematics education that most school students do not find these processes straightforward. Even when students seem to understand the function of proof in the mathematics classroom (see, Hanna, 1989; de Villiers, 1990), they still frequently fail to employ a method of reasoning about the truth of a conjecture that is mathematically acceptable (see, Fischbein, 1982; Healy and Hoyles, 2000).

There is a considerable corpus of knowledge in mathematics education that suggests theoretically how progress in reasoning might be characterized (see, Balacheff, 1988). There is, however, rather sparse description of the types of trajectories that progress in mathematical reasoning actually follows and how progress is (or is not) sustained over time. To begin to answer such questions, longitudinal data are needed, and to illuminate conceptual understanding, there is a need for these data to be derived from instruments specially designed for this purpose, alongside qualitative evidence to aid the interpretation and contextualisation of the results of quantitative analyses.

The Longitudinal Proof Project set out to track the progress in reasoning of a sample of Year 8 students (average age 13½ years) who were likely to be placed in top sets for mathematics two years later in Year 10. Data were collected through annual testing of students in highest-attaining classes from randomly selected schools within nine geographically diverse English regions. A total of 1512 students (named here proof-students) from 54 schools (111 classes) completed all three tests[i].

A proof test was designed and piloted for each annual survey comprising items in number/algebra and geometry, some in open format and some multiple-choice. At the outset of the project, a theoretical framework for the tests was articulated based on previous research (see for example Healy & Hoyles (2000) and Hoyles & Healy (in press) and work cited in these papers): for example, the framework included ascertaining whether or not a mathematical explanation justifying an answer could be constructed, and (in geometry) distinguishing between reasoning from the basis of perception or from geometrical properties. The next steps in the design of the proof tests were to analyse the curriculum for each year group, write items that were appropriate to the curriculum (as well as fitting the framework), and discuss the items with teachers in 5 pilot schools. Finally, the tests were piloted[ii] in order to assess whether the types of reasoning anticipated in the framework could be distinguished among the students' responses. Some items, called *core* items, were used in each of the annual tests. We then devised a classification scheme for students' responses to all the open response items, based on our framework and on insights gained from trialling[iii].

As well as the proof test, the Year 8 students completed a baseline mathematics test constructed from selected items from the Third International Mathematics and Science Study (Keys, Harris & Fernandes, 1996). Performance on this test provided a baseline measure of a student's general mathematical attainment against which to compare performance in reasoning. Raw scores on the statutory Key Stage 3 mathematics tests taken by all Year 9 students in England[iv] were gathered a year later and also used as an alternative baseline measure.

First, we illustrate how we identified and interpreted trends over time in students' responses to individual items in the proof tests, by reference to analyses of responses to one geometry item.

## Tracing trends in categorical data and exploring gender differences

G2b (shown in Figure 1) was a core item and non-standard (i.e. not familiar to students from school mathematics). G2b was also rather different from most other items in that the vast majority of students could give the correct answer (that the required fraction was a quarter), so the focus of analysis was on whether students could explain their correct solution (a more difficult task) and on whether their explanations 'improved' over time.

Squares A and B are identical. One corner of B is at the centre of A.

What fraction of A is overlapped by B ?

. . . . . . . . .

Explain your answer.

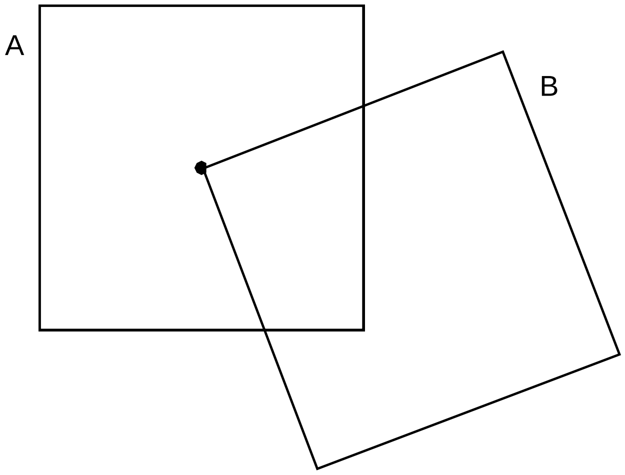
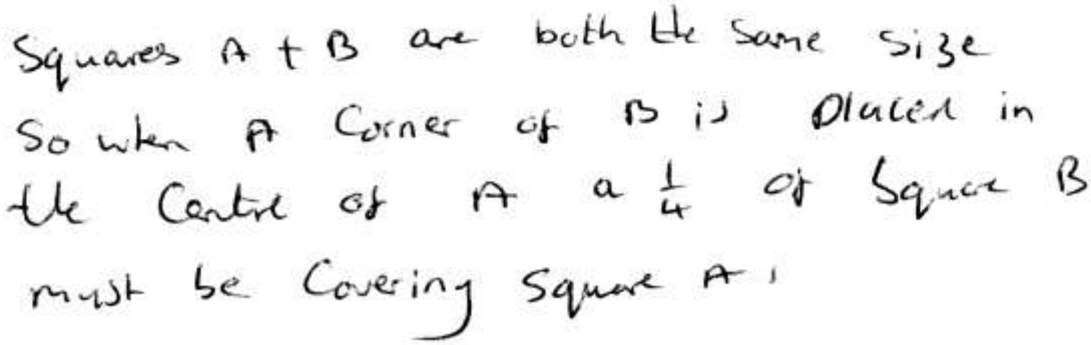


**Fig 1: Item G2b, a core item used in each proof test**

The classification scheme for the explanation part of G2b consisted of four broad categories, hierarchically-ordered according to our assessment of the quality of mathematical explanation. They ranged from category 1 (c1) (lowest level), where students gave no conceptual explanation (measured or produced a visual estimate), to c2, where explanations included reference to mathematical properties but without apparent deductive reasoning, to c3, where explanations were correct but lacked generality, and finally to c4, where explanations were judged as adequate for this age group[v]. We also scored the responses based on this hierarchy, for purposes of statistical analysis. A typical c2 response is illustrated in Figure 2, which consists essentially of a restatement of the givens.



Squares A + B are both the Same size
So when A Corner of B is placed in
the Centre of A a $\frac{1}{4}$ of Square B
must be Covering Square A.

**Fig 2. A typical category 2 response that simply restates 'given' properties**

Our analysis of the distribution of these response categories for G2b over the three years revealed that, according to our hierarchy, almost as large a proportion of students (25%) gave a lower level response in Year 10 than they had in Year 8 as gave a higher level response (26%) in Year 10 than in Year 8. A similar pattern occurred from Year 8 to Year 9 and from Year 9 to Year 10.

We were surprised by this seemingly dramatic lack of progress, which was unusual for our items, where responses tended to show steady development. We therefore questioned the validity of our original hierarchy for the G2b categories, from a student's perspective. Perhaps students held an alternative view about the nature of a 'good' mathematical explanation? To explore this conjecture we sought to validate the G2b classification hierarchy against a measure of a student's general mathematics achievement, by calculating for each year's data, the mean KS3 score achieved by the students whose responses fell into each of our categories. Surprisingly, given the fact that the KS3 score was a significant predictor of success in reasoning overall (see later), we found that the KS3 score means of the students who gave c2 responses were sometimes as high or even higher than the means of the students who gave responses in higher-level categories.

To help us to interpret this unexpected trend, we used qualitative methods in the form of analyses of individual student interviews in which we asked students whose responses in Year 10 were at a 'lower level' than in Years 8 or 9 to rank their responses and justify the ranking.

We provide an illustrative example of one such student, P, who gave a c3 response in Year 9 (in which she transformed the figure to a specific case that made the answer obvious) but gave the c2 response in Year 10 shown in Fig. 2, above. When asked which of the two responses she preferred, P chose the Year 10 response, since it "seems to prove it slightly more, in my mind …", and in particular because of "the part when a corner of one square is placed in the centre of the other". It seems that P preferred to re-present given properties rather than consider a specific, albeit illuminating case.

As a result of using these mixed methods, we were therefore able to identify an unexpected trend, the substantial number of students who gave c2 responses, and interpret it, as not because these students were unable to give 'higher level' responses, but because they valued c2 responses in ways we had not predicted: the responses displayed an apparent generality, and made reference to mathematical properties, even if these properties were only the ones given in the question and not derived by reasoning.

Another surprising trend identified through our quantitative analysis was that the KS3 mean score of students who omitted G2b or gave no explanation *increased* from Year 8 to Year 10. Again, we used student interviews to find out why, and these indicated that some students simply could not be bothered to give an explanation again, as illustrated in the following extract from an interview with a high-achieving boy:

> ...I remember this question. It's really clear that it is a quarter, but it's really hard to put into words why. So I've just written less and less each time, because I've done it before.

We therefore conjecture that recognition of repeat items was one factor in explaining the downward trend. Evidence from case studies also suggested that this was not the only reason, and some students exhibited more general 'test exhaustion', a phenomenon we first noticed when teachers and students in our case studies following the Year 9 test, complained that mathematics lessons had become dominated by preparation for the summer term's KS3 tests. We also noted that students' written reactions to the tests (that we requested each year) showed a marked increase in negative comments, as illustrated by the responses of the same student to the question "What did you feel about taking part in this survey?" in Years 8, 9 and 10:

> Year 8 response: I felt interested and quite pleased really.
>
> Year 9 response: I felt a little annoyed because no-one had told me what it was for and I detest unnecessary tests.
>
> Year 10 response: I feel that it is all right to do so – however we should be able to make a choice whether we want to spend nearly an hour of time that could be spent on doing useful things for exams on a test that won't affect our future or give us any qualifications.

To investigate gender differences in the distribution of categories of response to each item, the non-parametric Mann-Whitney U test (with many ties) was used to compare the mean ranks or median positions of boys and of girls (with the categorical scale treated as an ordered score). For G2b, this analysis showed a significant difference each year in responses in favour of girls ($p<0.02$, $p<0.002$, $p<0.005$ for Years 8, 9 and 10 respectively). That differences are significant is perhaps not entirely surprising, given the size of our sample, although it should be noted that we did not find significant gender differences on all our items and some differences favoured boys. However, the distribution of responses for girls and boys showed a larger percentage of girls than boys had indeed every year produced explanations in the highest level category suggesting that this difference was substantive and worthy of investigation. Since this consistent trend was only identified at the end of the study, it could not be followed up with student interviews or case studies.

In the next section, we describe how our analyses included investigation by mixed methods of factors that might affect students' development in reasoning that went beyond individual characteristics of attainment and gender.

## Identifying predictors of development and characterizing outliers

We devised two questionnaires: first to collect data on school factors, relating for example to school locality, student age range, the school's general achievement at GCSE[vi], and the policy and general practice of the school's mathematics department (text books used and examination boards followed); and second, to obtain data from the teachers of the classes of proof-students, on their qualifications, age, experience and courses attended. Students' total scores in algebra and in geometry were calculated for each yearly data set, and bivariate multilevel response models for the joint analysis of algebra and geometry were fitted to identify all the variables of different types (student, class, school) that might account for success in any one year in the proof test, and for any progress made. The first step in the modelling process was to fit a basic model to determine whether there was any detectable variation in students' responses at the school-, class- and student-levels in a model unadjusted for any explanatory variables.[vii] Then the most appropriate statistical bivariate model (see Woodhouse, 2002; Wang, 2002) was selected by the use of a forward and backward selection strategy (that is adding variables one at a time, testing the significance of the change and then reversing the process).

In each year's analyses, we found that the most significant predictors of proof performance in algebra and geometry were the student variables of baseline score[viii] and gender (in favour of girls), with few class- or school-level variables reaching significance. However, some variables were significant in the Year 10 analysis, including a teacher's years of experience, a professional development score (called *cpd*, see later) for geometry, and the % of proof-students who were to be entered for the GCSE higher tier (for algebra)[ix]. These factors helped us to characterise the schools we identified as positive outliers (see later) and to identify issues to investigate in our case studies of a sample of these outliers (for example, to find out about policies for selection of students to be entered for GCSE higher tier).

In Years 8 and 9 our case studies revealed some common characteristics, which we attempted to quantify for use in the modelling process in Year 10. For example, *all* the positive outlier schools visited were situated in relatively middle class, semi-rural areas of England, with catchment areas that were predominantly stable and white. This led us to question our assumption that the Year 8 baseline mathematics test or the Year 9 Key Stage 3 test would take up most of any variance in performance in reasoning due to social class. We therefore included, for the first time, in the Year 10 model the variable *fsm* that recorded the percentage, in each school, of students eligible for free school meals[x]. We hypothesised that the inclusion of *fsm* would allow the identification of factors that were specifically associated with developing mathematical reasoning, by taking account of higher performance related to other influences. In fact this was the case. When *fsm* was added to the models in Year 10, a significant negative effect on performance was noted and remained significant when baseline was used in conjunction with all the other significant explanatory variables (see models using KS3 score as baseline reported in footnote viii).

We noted a second common factor in some of the case studies, which concerned the teacher's engagement in 'extra' professional activity in mathematics. In Years 8 and 9, we had identified a range of factors to assess this engagement, including for example, numbers of days in course attendance and membership of subject associations. Since none of these factors individually proved to be significant, we combined them in Year 10 into a single linear measure, the *cpd* score, which did indeed prove to have a significant positive effect on geometry, but not on algebra, scores. We conjecture that this result might reflect a difference in emphasis on the two domains in the mathematics curriculum and the lack of experience in geometry among many, particularly younger, teachers (see Hoyles, 1997).

A third common factor identified from the Year 8 and Year 9 case studies was that the mathematics departments were largely stable, with a core of experienced staff who had been at the school for many years, alongside some ‘new blood’ (a new head of mathematics or new young teacher). In order to assess this stability factor, prior to the Year 10 analysis, we collected relevant data from all our schools, and used it in our models. We were however unable to find any significant effects, indicating that either the variable was also a feature of less successful schools, or that it was simply too crude to capture the ‘stability-mixed-with-change’ characteristic we had noted.

Returning to multilevel analyses, we report that as well as identifying factors significant in students’ responses, this analysis also played a major role in identifying a sample of schools or classes to case study. We termed these ‘positive outliers’, in that their students performed significantly better on a proof test or made more progress in reasoning between years than would be predicted from the models (see Healy and Yang, 2003). We now describe the process of identification of these outliers and report some highlights of the case studies that ensued, in order to illustrate the power of combining methods.

First we obtained estimates of effects and variation *before* the addition of class and school variables that contributed to the 'better than expected' student performance, by using a model with only two predictors (student-sex and one baseline measure) in its fixed part[xi]. Since the basic model for the Year 10 analysis indicated significant variation only at class and individual levels, the model was collapsed to a two-level model with no estimation of school level variation (see footnote vii). Figure 3 presents an example plot of class-level residuals against their ranks, including error bars around the residuals corresponding to 1.96 standard deviations, obtained from analysis of Year 10 responses. The class residuals (indicated as triangles) can be interpreted as representing the amount by which the class mean deviated from the mean predicted in the model. An error bar completely above the dotted line thus corresponded to a class that was performing above the mean predicted by the model, with 95% confidence, and provided a means to identify outstanding classes[xii]. It is these classes that we called positive outliers.

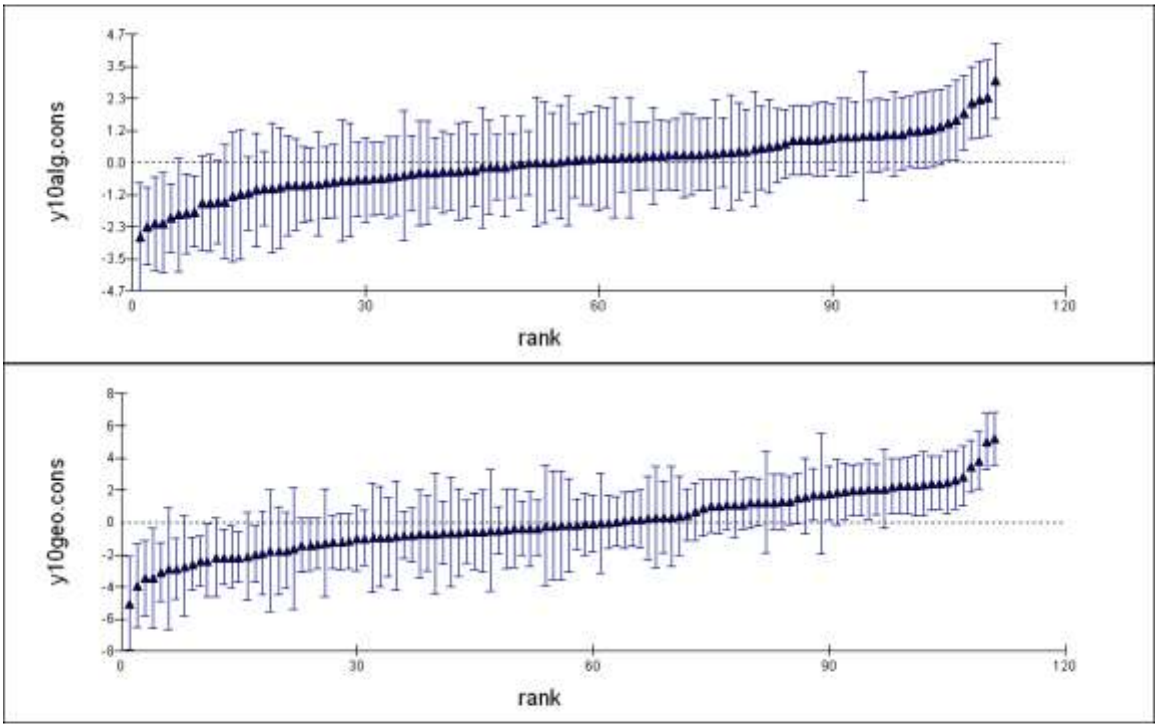


**Fig 3: Class residual intervals and error bars for scores in algebra (top) and geometry**

For each of the models constructed for the analysis of a particular year's data, we listed the schools in which one or more classes had been identified as positive outliers. There were classes in quite a number of schools in which students performed above the predicted mean in either algebra or in geometry, but only a handful that came up on all or nearly all the analyses, and each year we selected a small number of these for case study.

We now describe how we selected the case study schools following the Year 10 data collection. We undertook 16 basic analyses; four different base scores (Year 8 Baseline Mathematics Test score, Year 9 KS3 Test score, Year 8 Proof score and Year 9 proof score) were used for both algebra and geometry, each time with or without the variable *fsm*. Schools often fell out of any analysis that included *fsm*, but also there were schools in which one or more classes were positive outliers on many of the *fsm* analyses, but on few of the non-*fsm* analyses.

In the end we chose to case study five schools, A, B, C, X and Y. These schools were selected for slightly different reasons: A was represented in the list of positive outlier classes in almost all of the models constructed; classes from schools B and C appeared for most of the models and had been visited in previous years; and in schools X and Y positive outlier classes appeared *only* when models included the variable *fsm*. We would have wished to case study other schools, that displayed interesting gender difference or exceptional performance in one domain, but were unable to do so, due to limited resources.

We briefly describe the case studies of the chosen schools to indicate the range of data that came to light in our quest to interpret exceptional performance. All of the five case study schools were financially secure, either from good enrolment or from specialist school status, or both. All the mathematics departments, though they were stable, had interesting, possibly 'destabilising' influences, such as the arrival of a dynamic young second in the department, the adoption of specialist status in technology or the intervention of a maths advisor (a characteristic noted in previous years' case studies). All schools streamed in mathematics (as most schools in England), but notably all had a *superset* (in which the proof-students were situated) selected for and recognised as especially high achieving in mathematics. All or nearly all the students in these supersets were white, sometimes but not always because the school population was largely white. An experienced mathematics teacher using a mix of traditional and innovative ideas taught all the supersets. The criteria for superset status varied between schools: for example, in A, the superset students took mathematics GCSE a year early, whereas in X and Y (the schools that only became outliers after the inclusion of *fsm* in the models), the superset was the *only* set from which students were entered for the higher tier GCSE. Our case studies indicated that while the students in the supersets tended to be competitive as to who would achieve top marks, the teacher had managed to generate a collaborative classroom climate where students helped each other.

As well as identifying these factors in common among the case studies, what is also notable was that each case study revealed *unique* local factors that also helped to explain the positive outlier status. The most notable example was the existence nearby A of a research establishment which employed a huge number of science or mathematics PhDs (male and female), who were keen for their children to do well in general, but *particularly in mathematics*. Many of these parents sent their children to A, and exerted strong pressure on the school to excel in mathematics. They also could have had a more direct influence on their children's reasoning, as illustrated by an interview in which we asked a student whether a particular method that he had used on one of our items had been taught in class. He replied:

> No, but my dad is quite good at maths and in primary school he was helping me with this. We did quite a few of these towards the end of Year 5 and Year 6.

One further longitudinal quantitative analysis was undertaken in the last year of the project. A set of core items in algebra and in geometry had formed part of each proof test and 1512 proof-students had three scores on these items over the project period[xiii]. The total variance of the item scores could be attributed to 4 sources: school mean difference, class mean difference within school, individual students' differences within a class, and yearly differences within students. We decided to construct a series of multilevel repeated response models to explore quantitatively whether the students in general improved on the core items over the three years of the study and, if so, how this improvement varied across classes and individuals (see Healy and Wang, 2003). In particular we looked for gender differences in responses to core items over the three-year period and correlations at the individual and school/class levels. This analysis revealed that although there were no significant differences in the progress made by girls and boys on the core items from Year 8 to Year 10, there were significantly different patterns of progress: girls had a lower mean score than boys in Year 8, caught up with boys in Year 9 and then fell back slightly compared to boys in Year 10.

To investigate this further, we identified seven Year 10 classes where the students' scores had risen considerably in Year 9. By creating a variable that indicated 'being a girl' and 'membership of one of the seven classes', we estimated the mean progress of the girls from these special classes in Year 9, with interesting results: when it was added to the model, the significance of the variable that measured the improvement of all the girls' scores from Year 8 to Year 9 disappeared (whilst significance for other variables remained). We have difficulty in interpreting this result and have no qualitative data analysis to help us. It would have been interesting to investigate why girls in these seven classes made such marked progress, in fact large enough to account for all the differential progress in favour of girls. However, the very nature of this longitudinal statistical analysis meant that it *had* to be undertaken towards the end of a project, making case studies of classes to investigate a phenomenon that occurred 18 months previously (in Year 9) almost impossible.

By chance, we do have case study data for one of these 'special' classes, since the school happened to be a positive outlier in Year 8. Those data revealed that at the time of the Year 8 survey, the mathematics department had just changed from mixed ability teaching to setting, at the behest of a new head teacher. This was done hurriedly, with few appropriate materials and against the wishes of the mathematics department. This might explain why there was a marked improvement in proof scores in Year 9, when the new structure had settled down, but sheds no light on the 'special' progress of the girls in particular, as indicated by the analysis undertaken a year later.

## Concluding remarks

This paper has tried to pull out the strengths of mixing quantitative with qualitative methods in testing out ideas from different perspectives, problematising assumptions and identifying variables that might, with the use of just one method, have remained hidden or been accorded spurious significance. Tracing trends in categorical data provided insight into the development of mathematical reasoning, as illustrated by our analysis of responses to one item where the hierarchy of student explanations derived from theory was not completely supported by subsequent, quantitative analysis. The interpretation that this pointed to a mismatch between what teachers and students counted as good mathematical explanations, was largely derived from student interview analysis.

In other quantitative analyses of scores, we were able to gauge students' progress overall, and to identify significant predictors, some of which only came to light following longitudinal analysis. We were better able to appreciate the meaning of the predictors when contextualised through case studies, but reciprocally, attempting to quantify what we deemed were common factors from the case studies tested more rigorously our assumptions and perceptions. For example an initial assumption, concerning social class influences being 'wrapped up' in mathematics baseline, was not supported. In the last year, our case studies identified yet another common factor, the existence of a superset, but its robustness would need to be tested by adding it to later models.  However, it is important to report that not all influential factors revealed in case studies are quantifiable, such as the existence of unique local factors, which can only be 'uncovered' by qualitative means, and which appeared central to any explanation of exceptional performance or progress.

What is clear from this research is that any interpretation of statistical models requires not only detailed knowledge of the project design, but also of conditions in schools likely to influence performance in the area under study. When an effect is identified, it is difficult to know what led to the effect without detailed case studies. For example, the importance of entry to higher-tier GCSE mathematics in terms of raising expectations and motivation to reason mathematically, is plausible, but how is it manifested in practice? Similarly, we have given a plausible interpretation of the differences found between student performance in algebra and geometry, and the differential influence of a teacher's professional development in these two domains, but again, what is the effect on teaching? Confirming and critiquing interpretations and, more crucially, understanding how they are manifested in practice calls for more detailed classroom observation and study of teachers than was possible in our study. The challenge of such research is considerable, as it would not only be hugely time-consuming, but also hard to justify to schools, since it would require comparisons of negative with positive cases.

Findings from the mixed methods used in the Longitudinal Proof Project suggest that analyses of scores on tests (even those designed specifically to assess aspects of mathematics learning derived from theory) tend only to identify influential issues *not* directly related to teaching and learning mathematics (such as gender). While stating this, we also acknowledge, that with an even larger sample along with still more extensive analyses of the longitudinal data, we might have been more successful in pinpointing such issues, which then could have been included into statistical models. However, this would involve a design that would require much larger teams of personnel with different complementary expertise, working over several years. But even with the right design and appropriate staffing, we have noted above that it is hard to envisage how some investigations would be viable in the real world of schools. Moreover, given that it was only at the end of three years that some trends could be identified from longitudinal analysis, such research needs long term and sustained investment - and even then it might be impossible to use appropriate qualitative methods given the time lag.

We conclude by noting that combining methods promises attractive rewards as a consequence of bringing together a range of perspectives that separately and together offer unique insight into the complex process of learning in a subject domain. However the demands of this approach in time, effort and expertise must not be underestimated.

## Acknowledgement

We gratefully acknowledge the support of the Economic and Social Research Council (ESRC), Project number R000237777.

## REFERENCES

Balacheff, N. (1988). ‘Aspects of proof in pupils’ practice of school mathematics’. In D. Pimm (ed.), *Mathematics, teachers and children* (216-235). London: Hodder & Stoughton.

Brannen, J., (ed.) (1992). *Mixing Methods: Qualitative and Quantitative Research*. Aldershot: Avebury.

De Villiers, M. (1990). ‘The role and function of proof in mathematics’, *Pythagoras,* 24, 17-24.

Fischbein, E. (1982). ‘Intuition and proof’. *For the Learning of Mathematics*, 3 (2), 9-18.

Goldstein, H. (2000). *Multilevel Statistical Models.* 2nd Edition, London: Arnold

Goldstein, H. and Healy, M.J.R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society*, A **158**, 175-7.

Hanna, G. (1989). ‘Proofs that prove and proofs that explain’, *Proceedings of the 13th Conference of the International Group for the Psychology of Mathematics Education,* G.R. Didactique CNRS Paris, 45-51.

Healy, L. & Hoyles, C. (2000). ‘A Study of Proof conceptions in Algebra’, *Journal for Research in Mathematics Education*, 31 (4), 396-428.

Healy, L. & Yang, M. (2003). Multilevel Models of the Year 10 Data, in Küchemann, D. and Hoyles, C. (2003). *Year 8, Year 9 and Year 10 Technical Reports of the Longitudinal Proof Project*, http://www.ioe.ac.uk/proof/techreps.html.

Hoyles, C. (1997). 'The curricular shaping of students' approaches to proof'. *For the Learning of Mathematics*, 17(1), 7-16.

Hoyles, C., and Healy, L. (in press). 'Curriculum Change and Geometrical Reasoning'. In P. Boero (ed), *Theorems in School*. Dordrecht: Kluwer Academic Publishers.

Keys, W., Harris, S. and Fernandes, C (1996), *'Third International Mathematics and Science Study, First National Report Part 1'*. Windsor: NFER

Küchemann, D. and Hoyles, C. (2003). *Year 8, Year 9 and Year 10 Technical Reports of the Longitudinal Proof Project*, http://www.ioe.ac.uk/proof/techreps.html.

Wang, D. (2002). Multilevel Models of the Y9 Data, in Küchemann, D. and Hoyles, C. (2003). *Year 8, Year 9 and Year 10 Technical Reports of the Longitudinal Proof Project*, http://www.ioe.ac.uk/proof/techreps.html.

Woodhouse, G. (2002). Longitudinal Proof Project: Multilevel Models of the First Year's Data, in Küchemann, D. and Hoyles, C. (2003). *Year 8, Year 9 and Year 10 Technical Reports of the Longitudinal Proof Project,* http://www.ioe.ac.uk/proof/techreps.html.

[i] Initially 2,663 Year 8 students from 63 schools (114 classes), who were expected to be in the top set for mathematics when they reached Year 10, were surveyed in June 2000, using a specially designed Year 8 proof test. In June 2001, the same students (with some attrition) were tested again using the Year 9 proof test, which included some questions from the previous test together with some new or slightly modified questions. The same students (again with some attrition) were tested in the third year, in Year 10 in June 2002, with the same aims of testing reasoning and development in reasoning. Among the students who dropped out were those from nine schools, which left the study over the three year period, those who had moved to other schools or were absent on the day the test was administrated. Additionally, the design of the study was such that students who moved from the top mathematics set or band were not, in most cases re-tested. This meant that the sample included only the higher attaining students by the final year. A small number of schools adopted the practice of re-testing even those students who had moved to other classes and for this reason the number of classes involved in some schools increased over the study.

[ii] Using about 100 individual students and 3 whole classes within the 5 schools.

[iii] The resulting scheme was therefore not 'definitive' but depended in part on the particular responses we happened to sample during the development phase and on the particular characteristics that we

judged to be important. Students' written explanations were also often rather vague and cryptic, so categories inevitably depended on where we decided to draw boundaries.

[iv] Most of our students took the Level 6 - 8 tests, but where students took the Level 5 – 7 tests their scores were converted to the Level 6 – 8 equivalent, using a conversion table kindly supplied by the Qualifications and Curriculum Authority (QCA).

[v] A fifth category was used where the item was omitted or completely incorrect.

[vi] General Certificate of Secondary Education.

[vii] By the third year of the project, no significant differences in school-level variation were found, although at the class- and student- levels, significant differences in variation were associated with scores in both algebra and geometry. In the bivariate models constructed after the basic model, only two levels of variation were hence examined.

[viii] The number of baseline measures increased each year and by Year 10 different models were built with four different baseline measures (scores on Year 9 proof test, scores on Year 8 proof test, scores on the Year 9 Key Stage 3 test score and scores on the baseline mathematics test administered in Year 8).

[ix] The final models using KS3 score as baseline were:

Predicted algebra score in year 10 = 10.642 + 0.070*ks3* + 0.478*girl* – 0.127*fsm*+ 0.018*higher-tier* – 0.028*a to c* – 1.54*small-class*

Predicted geometry score in year 10 = 14.905 + 0.098*ks3* + 0.369*girl* – 0.121*fsm* + 0.350*cpd-tot* + 0.058*years teaching* – 2.805*small-class*

[x] In the previous year we had not collected any measure of a student's social class.

[xi] A third predictor, *fsm*, was added in Year 10.

[xii] This identification is conservative compared to the new criterion value proposed by, for example, Goldstein and Heath (1995).

[xiii] This is a typical dataset with repeated measures at student level, which leads to a 4-level structure of hierarchy as repeated tests (4536) nested within student, students (1512) nested within class, and classes (111) nested within schools (54).