INSTITUTE OF
EDUCATION
UNIVERSITY OF LONDON

Mathematical Sciences

Technical Report On the Nationwide Survey

Justifying and Proving in School Mathematics

Lulu Healy and Celia Hoyles

February 1998

Funded by the Economic and Social Research Council

ACKNOWLEDGEMENTS

First and foremost our thanks go to all the teachers and students who agreed to take part in the survey, to our colleagues in the mathematics education community in England and Wales who initiated contact with the schools and to the local fieldworkers who actually administered the survey.

We are grateful to all our colleagues in the Mathematical Sciences Group for their support and advice at all stages in the design, administration and analysis of the survey, and to the members of the project advisory group for their invaluable input at critical stages.

We are especially indebted to Min Yang for her statistical advice, and in particular for her work in constructing the multinomial models for the analysis of the categorical data. Thanks go to the other research officers who made particular contributions to the work in this report: Jill Bruce, Paul Clifford and Joyce Shaw. Thanks also to Dorothy Pattman for her patience and assistance in the production of this report.

Finally, we gratefully acknowledge the financial support of the Economic and Social Research Council (grant No. R000236178) who made this research possible.

CONTENTS

EXECUTIVE SUMMARY

Proof is at the heart of mathematical thinking, and deductive reasoning, which underpins the process of proving, exemplifies the distinction between mathematics and the empirical sciences. Developing the ability to recognise and construct chains of logical argument based upon agreed rules and procedures is fundamentally important for those aiming for careers that rely upon mathematical literacy. The process of building a valid proof is clearly a complex one: it involves sorting out what is given — the mathematical properties that are already known or can be assumed — from what is to be deduced, and then organising the transformations necessary to infer the second set of properties from the first into a coherent and complete sequence. Research in mathematics has consistently highlighted students' difficulties in engaging with formally-presented, analytical arguments and understanding how these differ from empirical evidence. The current National Curriculum for mathematics prescribes an approach to proving, maybe as a response to these student difficulties, in which the introduction of formal proofs is reserved for 'exceptional performance', and thus delayed until after students have progressed through early stages of reasoning empirically and explaining their conjectures. Most of the requirements to explain and justify take place within investigations driven by numerical data, as part of the attainment target, Using and Applying Mathematics.

The project, Justifying and Proving in School Mathematics[1], started in November 1995 with the aim to examine the impact of the National Curriculum on high-attaining Year 10 students' views of and competencies in mathematical proof. In particular, it set out to:

- describe the characteristics of mathematical justification and proof recognised by high-attaining Year 10 students;
- analyse how students construct proofs;
- investigate the reasons behind students' judgements of proofs, their performance in proof construction and their methods of constructing proofs.

Two questionnaires were designed, piloted and refined, a student proof questionnaire and a school questionnaire. The proof questionnaire comprised a question to ascertain a student's views on the role of proof, followed by items in two domains of mathematics — arithmetic/algebra and geometry — presented in open and multiple-choice formats. In the former format, students were asked to construct one familiar and one unfamiliar proof in each domain. In the latter format, students were required to choose from a range of arguments in support of or refuting a conjecture in accordance with two criteria: which argument would be nearest to their own approach if asked to prove the given statement, and which did they believe would receive the best mark. The school questionnaire was designed to obtain data about a school and about the mathematics teacher of the class selected to complete the proof questionnaire. These teachers also completed all the multiple-choice questions in the proof questionnaire, to obtain their choices of argument

1 Funded by ESRC, Project Number R000236178

and to identify the proof they thought their students would believe would receive the best mark.

After piloting with 182 students, the survey was administered to 2,459 Year 10 students (14 or 15 years-old) from 94 classes in 90 schools. All the students were in top mathematics sets or chosen as high-attaining by the mathematics departments. Key Stage 3 test scores[2] of the students who completed the questionnaire were provided by the schools and these ranged from Level 5 upwards with an average of 6.56. The schools were spread across England and Wales and included mixed and single sex schools in different locations (urban, rural, suburban), operating diverse forms of selection procedures on entry.

The following student outputs were analysed: scores on the four constructed proofs and the forms of argument used; scores for assessing the correctness and generality of the arguments presented in the multiple-choice questions (one score for algebra and one for geometry); choices in the multiple-choice questions; and views of the role of proof. Descriptive statistics were used to describe patterns in student response, followed by multilevel modelling using data from the school questionnaire to identify factors associated with performance and how these varied between schools. In presenting these findings, we have chosen to interpret some associations causally, while recognising that these interpretations must be treated with caution.

1. High-attaining Year 10 students show a consistent pattern of poor performance in constructing proofs.

Overall, the performance of the students in the constructed proof questions was very disappointing. The average score for a constructed proof was less than 1.5 (half of the maximum score) and for the unfamiliar questions, well below 1. It is important to stress that a proof was scored merely on the basis of the correctness of the argument, and not on its presentation. Many students were unable even to begin to construct a proof (between 14% and 62% scored 0) and, if they did make a start, between 28% and 56% could only indicate relevant information unconnected by logical argument, thus scoring only 1. The percentage of students showing evidence of deductive reasoning varied according to the mathematical content of the question, with rather more, 40%, in the familiar algebra proof and very few in the harder questions in both domains (about 10% of students). Empirical verification was the most popular form of argumentation used by students in their attempts to construct proofs, and in problems where empirical examples were not easily generated, the majority of students were unable to engage in the process of proving.

Among those students who did not rely exclusively upon empirical evidence, arguments written in narrative form were more common than formal presentations, and in general were associated with a higher incidence of deductive reasoning. However, in the case of the unfamiliar geometry proof, it was amongst the tiny group of students, (about 10%),

2 National tests administered to all students, aged 13-14 years, in England and Wales at the end of Year 9 of the National Curriculum. Average level of attainment at this stage is between levels 5 and 6.

who attempted to construct a formal proof, that the highest proportion of deductive arguments was found (54%).

2. **Students' performance is considerably better in algebra than in geometry in both constructing and evaluating proofs.**

Although there was little difference between domains in the (very small) percentage of students able to produce a completely correct proof, the picture of proof construction was rather more positive in algebra than in geometry, in that students seemed better able in algebra to identify the relevant mathematics and begin to construct logical arguments. When trying to prove the familiar conjecture — that the sum of two odd number is always even — 40% of students used some deductive reasoning, whereas only 24% used any deduction when proving the equally familiar statement — that the sum of the angles of a quadrilateral add up to 360^{o}. Where the content was less familiar, a larger proportion (56%) of students was able to isolate relevant pieces of knowledge in algebra as compared to only 28% in geometry, where 62% did not know where to begin.

Students were also considerably better in algebra than in geometry at assessing whether an argument was correct and whether it held for all or only some cases within its domain of validity. In fact, not one student amongst the sample was able to assess correctly the ten geometry arguments, while 20 managed to do this in algebra.

3. **Most students appreciate the generality of a valid proof.**

Despite difficulties in evaluating particular arguments, the majority of students were aware that, once a statement had been proved, no further work was necessary to check if it applied to a particular subset of cases within its domain of validity. In contrast to all other questions, more students answered this question correctly in geometry (84%) than in algebra (62%).

4. **Students are better at choosing a valid mathematical argument than constructing one, although their choices are influenced by factors other than correctness, such as whether they believe the argument to be general and explanatory and whether it is written in a formal way.**

Significantly more students were able to select a correct proof than to write one. However, they were influenced by their view of the generality of an argument and how far they judged it to be convincing: they were more likely to choose arguments that they believed to be general and which they found helpful in clarifying and explaining the mathematics in question.

Students were also likely to make rather different selections depending on the two criteria for choice: the argument which most closely resembled the approach they would adopt or the one they believed would receive the best mark. For best mark, formal presentation was chosen frequently and empirical argument infrequently – even when the latter would have provided a perfectly adequate refutation. Such empirical counter-examples were much more common when students were selecting the refutation closest to their own approach. In algebra, students were less likely to choose an empirical argument than to construct one, and in all three multiple-choice questions, arguments presented in a prose-style were the most popular choice for own approach, with symbolic-algebraic forms the least popular. In geometry, patterns of choice for own approach were less clear-cut, although formal arguments were selected more frequently than in algebra.

5. **General mathematical attainment has a consistent influence on students' views of proof and their competencies in proving, although it is never the only significant variable associated with performance.**

Of all the factors associated with student responses, students' general mathematics attainment as measured by Key Stage 3 test score exerted the most consistent effect. In both domains, students with high, as compared to lower, Key Stage 3 scores constructed better proofs, were less likely to rely upon empirical evidence in their constructions and selections, and were better at evaluating arguments in terms of correctness and generality.

However, Key Stage 3 test score was never the only significant variable associated with student performance in proving, and other student factors, along with a range of particular characteristics of school and curriculum, were also found to be influential.

6. **Students' views of proof and its purposes account for differences in their responses.**

Students' views of proof and its purposes were associated with performance in a variety of ways: students with little or no sense of proof (over one quarter of the sample) were more likely to choose empirical arguments; those who recognised the generality of proof and its role in establishing the truth of a statement (over half of the sample) were better at constructing proofs and evaluating particular arguments; and in algebra, students who believed that a proof should be explanatory (over one third of the sample) were less likely than others to try to construct formal proofs and more likely to present arguments in a narrative form.

7. **In algebra, girls and boys perform significantly differently, with girls constructing better proofs than boys and choosing different forms of argument.**

In algebra, girls and boys performed significantly differently, with girls scoring higher in their constructed proofs. Girls also showed preferences for different forms of argument than boys, both in their own proofs and the arguments they chose, although there was no obvious pattern to these differences. Boys also appeared more susceptible to school influences, with scores on the familiar algebra proof varying according to the school attended, whereas girls' performances were similar across all schools.

8. **Teacher characteristics are not associated with students' competencies in proving.**

There was no variation in student response according to teacher variables, such as qualifications, sex and teaching experience, although it must be noted that almost all of the teachers in the sample were well-qualified mathematically. Neither did the teachers' responses to the survey in terms of their own choice of approach or their predictions of their students' choices for best mark appear to influence student response.

9. A range of school and curriculum factors are associated with performance.

School and curriculum factors influenced students' competencies in proving, although no one factor had an effect across all questions, even in the same mathematical domain. However, some general trends are identifiable.

(a) Students in classes with a larger proportion taking the higher- rather than the middle-tier GCSE paper are better at both constructing and evaluating proofs.

One factor consistently influential in student proof constructions and evaluations was the percentage of students expected to sit the higher-tier GCSE[3] paper: those from classes with a large percentage of students expected to sit this paper were likely to be better at constructing proofs and evaluating arguments in terms of their correctness and generality than similar students from classes where more would be entered for the middle-tier paper.

(b) Curriculum factors, such as the number of hours of mathematics teaching each week, and the textbook or examination syllabus followed, exert significant influences on student response in both domains, but are particularly apparent in algebra.

Curriculum factors were significant variables in student performance: in particular, the hours of mathematics teaching each week, the textbook used and examination syllabus followed affected the responses students made, especially in the multiple-choice questions. For example, more mathematics teaching reduced the likelihood of students choosing an empirical argument. These influences were all more apparent in algebra than in geometry.

(c) A specific emphasis on proof improves student performance.

In response to some questions, specific emphases on teaching proof appeared to improve students' performances: students in classes expected to write formal geometry proofs were more likely to be able to do so, and those from classes where proof was taught as a separate topic were better than others at evaluating arguments in algebra.

---

3 The GCSE is the public examination taken by students in England and Wales at the end of their compulsory schooling (age 16 years). Students are entered to one of three levels in the examination, the foundation-, middle- or higher tier. Although there is overlap in the grades obtainable from taking the different tiers, there are ceiling grades for the lower tiers.

10. After taking into account all the factors found to influence student performance, there remains unexplained variation in the responses of students attending particular schools.

Although there was more unexplained variation in performance within schools than between schools, school variation was found, and outlier schools identified whose students performed significantly better or worse than predicted on more than half of the scores. School variation was more evident in geometry than in algebra, with schools in the former case differentially affecting students' choices for their own approach, the forms of argument students used in constructed proofs, student preferences for formal arguments and student ability to assess the correctness and generality of an argument.

11. Summary and conclusions.

The major finding of the project is that most high-attaining Year 10 students after following the National Curriculum for 6 years are unable to distinguish and describe mathematical properties relevant to a proof and use deductive reasoning in their arguments. Most are inclined to rely upon empirical verification. However, students perform more successfully when it comes to choosing rather than constructing correct proofs. The majority also recognise that a valid proof is general and accord high status to formally-presented arguments, even while valuing arguments that convince and explain.

The research indicates that the ability to construct, assess or choose a valid proof is not simply a matter of general mathematical attainment. Clearly this has an influence, but at least some of the poor performance in proof of our highest-attaining students may simply be explained by their lack of familiarity with the process of proving. Far too many students have little idea of this process and no sense of proof, which, our findings suggest, can hinder their ability to construct and correctly evaluate proofs.

The study was unable to identify teacher characteristics associated with different student responses, although school and curriculum factors did prove to be influential. Student performance in geometry was consistently poor and is a major cause for concern. This again, we suggest, is a matter of curriculum emphasis. The high-attaining students in our survey had little familiarity with geometrical structures and relationships, even of the simplest kind, and were certainly unused to explaining geometrical phenomena. It could also be argued that the fact that school effects were more apparent in geometry than in algebra was a result of their relative emphases in the curriculum – as so little is prescribed in geometry, more leeway is available for some teachers to make quite a difference if their situation makes it possible and they so decide.

In contrast to the absence of any curriculum requirement to engage in geometrical argument, students, under our existing guidelines, gain plenty of experience in constructing empirical verifications and refutations. They are accustomed to number/algebra investigations, where results have to be presented and explained, but where the focus of explanation appears to be less on the mathematical properties and relationships which underpin constructions than on the output data. The research suggests however, that many students do come to value general and explanatory arguments through these investigative activities, but this fertile ground is not exploited to introduce mathematical proof and face students with the challenge of setting out a mathematical argument in a coherent and logical manner.

Particular curriculum influences on student responses were apparent in the survey, although generally their effects varied from question to question, suggesting that familiarity with mathematical content rather than general competencies in proving was the dominant influence. Nonetheless, the study does identify influential factors which suggest that more challenge and more attention to proving could enhance performance: students in classes with a larger proportion taking the higher- rather than the middle-tier GCSE paper, or where proof is explicitly addressed and the writing of formal proofs encouraged, do better than their counterparts in other classes.

Taken together the results of our study suggest that, in the forthcoming review of the National Curriculum for mathematics, attention should be paid to the coverage of geometry and more generally to the approach to proof. We suggest that more explicit efforts should be made to engage students with proof while discussing with them the idea of proof at a meta-level, in terms of its meaning, generality and purposes. This would involve finding ways of balancing the need to produce a coherent and logical argument with the need to provide one that explains, communicates and convinces. This implies that alongside the curriculum emphases on measurement, calculation and the production of specific (usually numerical) results, more consideration should be given to appreciating mathematical structures and properties, the vocabulary to describe them, and simple inferences that can be made from them. Our evidence suggests that students could well respond positively to the challenge of attempting more rigorous and formal proofs alongside informal argumentation, and that developing approaches where this might be accomplished in the context of geometry as well as of algebra, would be a useful way forward.

We report on the results of the paper and pencil survey administered during phase 1 of the research project, Justifying and Proving in School Mathematics.

1. AIMS

The aims of the survey were:

- to describe the characteristics of mathematical justification and proof recognised by high-attaining Year 10 students;
- to analyse how students construct proofs;
- to investigate the reasons behind students' judgements of proofs and their methods of constructing proofs.

2. THE DATA

Two questionnaires were designed, a student *proof questionnaire* and a *school questionnaire*. The *proof questionnaire*, targeting high-attaining Year 10 students, comprised questions in two domains of mathematics — arithmetic/algebra (A) and geometry (G). It was pre-piloted through interviews with 68 students in 4 schools to test whether it was pitched at an appropriate level and was sufficiently engaging for students. Following the pre-pilot, items were removed if too easy or modified if too hard.

In order to be able to make comparisons between responses in algebra and in geometry, the format in each domain followed an identical pattern. Additionally, the order of questions in any one domain was such that information from earlier questions could be used later.

The following points summarise the final structure of the questionnaire:

- A question to ascertain a student's views on the *role of proof*: "What is proof for?"
- Six *multiple-choice* questions, (3 in each domain), where students choose from a range of arguments according to two criteria: what they would do if asked to prove the given statement and which of the proofs they believe would receive the best mark. Some of the proofs or refutations are correct and some incorrect. They are presented in a *variety of forms* – empirical, exhaustive, visual, narrative and formal.
- Direct proofs with *familiar* mathematical content in the first multiple-choice questions (A1, G1)[2], a *false conjecture* in the second (A5, G5) and direct proofs with *less familiar* mathematical content in the third (A6, G6).

2 A3, G3 presented visual proofs of the same statement given in A1, G1.

- Student evaluations of the *generality of each argument* presented in A1, G1, A6 and G6, and their assessments of how far they feel each one *explains* and *convinces* them of the validity of the given statement.

- Questions to ascertain a student's assessment of the *generality of a valid proof* (A2, G2).

- Four questions (2 in each domain) which require the students to *construct a proof*, with the first question (A4, G4) more familiar than the second (A7, G7).

As mentioned above, as part of the first and last multiple-choice questions, students were asked to assess the generality of each of the arguments presented. The correctness of their evaluations was scored by what is called a student's *Validity Rating* (VR) for the argument. Students were also asked to assess how far each argument explained the proof and convinced them of its truth, and these assessments were combined to give a score, called its *Explanatory Power* (EP).

Simultaneously with the development of the student proof questionnaire, a *school questionnaire* was designed to obtain information about a school — the type of school, its selection and setting procedures, the hours spent on mathematics, the textbooks adopted and the examinations entered. We also sought *data from the mathematics teacher* of the class selected to complete the proof questionnaire, to provide information on his or her background, qualifications, reactions to the place of proof in the National Curriculum, the approaches adopted to proof and the proving process in the classroom and the percentage of the class who would be entered for the GCSE higher tier. These teachers were also expected to complete all the multiple-choice questions in the student proof questionnaire, in order to ascertain their choice of proof and which proof they thought their students would believe would receive the best mark. Finally, the Key Stage 3 test scores of all the students who completed our questionnaire were provided by the schools.

Both the questionnaires were piloted with 182 students in 8 schools. The final versions of each are available in Appendices 1 and 2.

SAMPLE AND ADMINISTRATION

The survey was administered to 2,459 students from 94 classes in 90 schools with the 94 class mathematics teachers completing the proof questionnaire and school questionnaire. The schools were spread across England and Wales, 29 in urban, 25 in rural and 36 in suburban settings. 65 were LEA funded, 18 Grant maintained and 7 were Church schools[3]. 81 schools operated no form of academic selection, 7 selected all of their intake on a academic basis and the remaining 2 operated some form of academic selection. The majority of the schools (77) were mixed-sex, with 9 girls' schools and 4 boys' schools.

From the school questionnaire, we obtained data about our sample to give more detail about the context of mathematics teaching. Nearly all of the classes who took part in the

3 Independent schools were not included since they are not obliged to follow the National Curriculum for Mathematics.

survey were the top mathematics set, with only 3 second sets and 1 mixed ability. Of the schools in which students were set by ability, 33 classes (35%) had been set in Year 8, 30 (32%) in Year 7 on entry to the school, 26 (28%) in Year 9, and 4 (4%) in Year 10. The average number of hours of mathematics teaching per week was 3 hours with a range from 2 to 6 hours with the majority of classes (87) receiving between 2.5 and 3.5 hours teaching. The proportion of students within each class expected to be entered for the GCSE higher-tier paper varied from 0 to 100%, with an average across all classes of 80%. The most popular examination syllabi followed were SEG[4] (33 classes, 35%), MEG[5] (26 classes, 28%) and London (22 classes, 23%), leaving only 13 classes, (14%) adopting other syllabi. By far the most common textbook or scheme used was SMP[6] (32 classes, 34%), with 12 classes (13%) using books by Vickers, 11 (12%) by Holderness, 9 (10%) by Rayner. The remaining 30 classes (32%) followed a total of a further 10 different schemes or textbooks.

The most common approach to the teaching of mathematical justification was through investigations (72 classes, 77%), with only 16 classes (17%) addressing it as a topic area in its own right. The remaining 6 classes (6%) used neither of these approaches, and further details about the teaching of mathematical justifications were not specified. Students were more likely to be expected to read and write proofs in algebra than in geometry: students in 62 classes (66%) were expected to read algebraic proofs compared to 41 classes (44%) for geometric proofs; writing algebra proofs was expected in 48 classes (51%) but only 30 classes (32%) for geometry.

On the whole, the students in our sample were taught by well-qualified teachers with 89 (95%) holding a good qualification, 4 (4%) holding one that was acceptable and only 1 (1%) with a poor qualification[7]. The teachers were evenly split in terms of gender, 48 (51%) female and 46 (49%) male with years of teaching experience ranging from 1 to 30 years. Most of the teachers felt that the emphases on mathematical justification and formal proof in the current National Curriculum for Mathematics (DfE, 1994) was about right — 53 teachers (56%) in the case of mathematical justification and 57 teachers (61%) for formal proof). Very few teachers felt that either one was over-emphasised, with only 4 teachers (4%) holding this view with respect to mathematical justification and 2 teachers (2%) for formal proof. In contrast, a substantial minority felt both aspects were under-emphasised, 37 teachers (39%) for mathematical justification and 34 teachers (36%) for formal proof.

The sample of 2,459 students was made up of 1305 girls and 1154 boys, with a mean Key Stage 3 score of 6.56 (1 level 4, 133 level 5, 920 level 6, 1109 level 7, 162 level 8 and 133 unknown).

The questionnaires were administered between May and July 1996 by local fieldworkers who received a detailed set of instructions specifying administration procedures (see Appendix 3). This mode of organisation ensured consistency in administration and a 100% return of student questionnaires. It also had the benefit of guaranteeing the collection of teacher and school data, since while the students answered their

4 Southern Examinations Group.
5 Midlands Examinations Group.
6 School Mathematics Project.
7 The classification specified in the Cockcroft report (DES 1982).

questionnaires, their teacher filled in the relevant parts of the proof questionnaire and completed the school questionnaire.

CODING STUDENTS' RESPONSES

Schemes for coding the questionnaires were devised and, between June and August 1996, all the survey scripts and the school questions were coded and the data stored electronically. Three researchers coded the same set of 30 survey scripts to assess inter-coder reliability.

A total of 190 variables were entered for each student proof questionnaire and 95 variables for a school questionnaire. An initial data entry error was calculated to be 0.00205 (2 mistakes every thousand entries). This was reduced by a second researcher checking every tenth script entered.

The coding schemes for the three types of question (role of proof, multiple-choice, proof constructions) and the inter-coder reliabilities are given followed by the methods for calculating the VR and EP scores.

a. Classification of student responses to the question "what is proof for?"

Student response	Code
Not answered	0
Answers relating to verification/"truth"	1
Answers relating to explanations, reasons	2
Answers relating to providing evidence	3
Answers relating to communicating to others	4
Answers relating to discovering new theories/ideas	5
Answers relating to ability/achievement	6
Answers relating to general validity, completeness	7
Answers including some reference to logical thinking	8
Other	9

Table 1: Coding scheme for role of proof (inter-coder reliability: 0.962; 0.938; 0.942)[8]

As some codes appeared very infrequently, initial coding of data was subsequently simplified into four categories: *Truth* (codes 1, 3, 7 and 8); *Discovery* (code 5); *Explanation* (codes 2 and 4); and *Other/none* (codes 0, 6 and 9).

b. Forms of proof and their correctness in the multiple-choice questions

Table 2 shows the codes used to distinguish the different forms of proofs and whether the proofs given in these forms in the multiple-choice questions were correct or incorrect.

8 Coding was undertaken by three researchers, three correlations are reported to show the pairwise correspondences.

Form of proof	Code	Correctness in multiple-choice
Empirical: Unelaborated calculations or measurements	1	Incorrect
Exhaustive: All possible cases tested	2	Correct
Enactive: Unelaborated description of actions and observations	3	Incorrect
Naive: Restatement of givens; statements of unhelpful or wrong "facts"	4	Incorrect
Analytical Formal (correct): Logical argument in formal mathematical language	5	Correct
Analytical Formal (incorrect): Incorrect, incomplete or illogical argument in formal mathematical language	6	Incorrect
Analytical Narrative: Logical argument, not in symbolic form	7	Correct
Visual: Diagram with visual clues showing the logic of the proof	8	Correct
Counter-example: Production of a counter-example with no elaboration	9	Correct

Table 2: Coding scheme for forms of proof (inter-coder reliability: 0.931; 0.949; 0.930).

c. *Classification of students' constructed proofs.*

Students' constructed proofs were classified according to two criteria: form of argument using the categories presented above in Table 2; and a score for correctness based on the codes shown in Table 3.

Evaluation of Constructed Proof	Code
No basis for the construction of a correct proof	0
No deductions but relevant information presented	1
Partial proof, including all information needed but omitting some steps of reasoning	2
Complete proof	3

Table 3: Coding scheme to evaluate constructed proofs (inter-coder reliability: 0.925; 0.954; 0.936).

d. *Generality of Proofs, Validity Rating, Validity Score and Explanatory Power*

How far students recognised the generality of a valid proof was assessed by one question in algebra, A2, and one in geometry, G2, that followed the first multiple-choice question in each domain. In both questions, students were asked: if a statement[9] had already been proved over a particular domain, whether any additional work would be required to prove if it held for a given subset of the domain. These questions were intended to distinguish those who correctly assessed the generality of a valid proof from those who did not. Students scored 1 if they knew that the proof was general and 0 if they thought it to be specific.

9 The statement 'proved' in the preceding multiple-choice question.

A student's validity rating (VR) was calculated for each argument presented in the multiple-choice questions A1, A6, G1 and G6. The VR was a score of 0, 1 or 2 based on students' responses to the following 3 questions about the argument:

- It had a mistake
- It showed the statement was always true
- It showed the statement was true for some examples.

An entirely correct profile of responses for any given argument scored 2; a profile in which the student correctly noted if the argument was general, specific or wrong but was unsure of other factors obtained a rating of 1; all other profiles scored 0.

The VR's of the six proofs in A1 and the VR's of the 4 proofs in A6 were combined to give an overall *validity score* (AVS) for algebra with a range 0 to 20. Similarly, a *validity score* (GVS) was calculated from the VR's of the proofs in G1 and G6, again with a range of 0 to 20.

The rating of the explanatory power (EP) was also calculated for each argument presented in the multiple-choice questions A1, A6, G1 and G6. This was based on students' responses to the following two questions about the argument:

- It showed why the statement is true
- It was an easy way to explain to someone who was unsure.

If students agreed with both statements, their EP for that argument received a score of 2; if they agreed with one or other of the statements the EP scored 1; otherwise the EP was 0.

ANALYSIS OF THE DATA

The main purposes of the analyses of the data from the proof questionnaire were fourfold: to describe students' views of the role of proof before undertaking the survey, to categorise their choices and assessments of different arguments and to score their constructions and evaluations of proofs, to establish factors associated with all these responses, and finally, to examine how these factors varied between schools by reference to data from the school questionnaire. To achieve these goals, descriptive statistics based on frequency tables, simple correlations and tests of significance were produced, followed by a more sophisticated modelling of the factors associated with student response. The survey was administered to whole classes of students with the same teacher, who therefore shared experiences that would be expected to lead to correlations in scores. A multilevel analysis (see Goldstein, 1995) of the dataset was therefore used. This had a two-level structure since there was one class per school in 85 out of 89 schools[10], so class and teacher factors were considered alongside school and curriculum factors as Level 2 variables.

10 One class was excluded from the multilevel analysis as no Key Stage 3 test scores were available.

Models were constructed for the 16 output measures (10 choices and 6 scores) shown in Table 4.

From multiple-choice questions	
Students' choice for their own approach	In algebra A1, A5, A6 In geometry G1, G5, G6
From constructed proof questions	
Constructed proof scores	In algebra A4, A7 In geometry G4, G7
Form of constructed proof	In algebra A4, A7 In geometry G4, G7
Algebra validity score (AVS)	Based on proofs in A1, A6
Geometry validity score (GVS)	Based on proofs in G1, G6

Table 4: List of output measures.

Codes derived from the two questionnaires that rarely appeared were eliminated and those remaining comprised the 34 input variables tested in the models of the 16 output measures. These are shown in Table 5 where they are grouped into categories according to their level: Level 1 (student factors) and Level 2 (school, curriculum, teacher and teaching factors).

Level 1 Variables: Student factors	
Views of the role of proof	Truth
	Discovery
	Explanation
Student characteristics	Sex
	Age
	Key Stage 3 test score (KS3 test)
Responses to questionnaire	Best mark[11]
	Proof as general (algebra)
	Proof as general (geometry)
	Validity ratings (VR)[11]
	Explanatory power (EP)[11]
Level 2 Variables:	
School factors	Location
	School sex
	Size of Year 10
	School selection procedures[12]
	Year students are set
	Number of sets
Curriculum factors	% GCSE higher tier
	Examination syllabus
	Main textbook/scheme
	Hours of mathematics teaching per week
Approaches to teaching proof	Through investigations
	As a separate topic
	Read geometry proofs
	Write geometry proofs
	Read algebra proofs
	Write algebra proofs
Teacher characteristics	Sex
	Years of teaching experience
	Qualifications
Teachers' views of the National Curriculum	Emphasis on mathematical justification
	Emphasis on formal proof
Teachers' responses to multiple-choice questions	Teachers' choice for their own approach[11]
	Prediction of student choice for best mark[11]

Table 5: List of input variables.

Findings are reported in the following twelve sections. First, we report on students' responses to the question *"what is proof for?"*. We then consider in some detail students' responses to the multiple-choice questions, the extent to which they appreciated the generality of valid proofs, and their attempts to construct proofs of their own. Next,

11 These variables occurred in the multiple-choice questions only.

12 Schools were split into two groups: those with no selection procedures and those who operated any form of selection. The category of selective schools therefore included church schools whose intake was chosen on religious grounds, as well as those with some form of academic selection.

we report on how well they were able to assess the generality or scope of validity of the arguments presented to them. In each case, we begin by describing patterns of response, after which we consider whether student, school, curriculum, teaching and teacher factors were associated with different response patterns and how these patterns varied between schools. Finally, we consider briefly teacher and school effects which are not well explained by the statistical models constructed.

The overwhelming first impression of the data is one of diversity in response both between and within the two mathematical domains. Nonetheless, despite this diversity, a considerable number of findings describing student choice patterns and constructions can be identified and their generalisability tested. Some findings hold for both domains, while others are specific to algebra or to geometry although comparisons can be made between 'equivalent' questions in each domain — that is between the 'easy' constructed proofs, or the questions in which false conjectures are presented to be 'proved' or refuted. In all the findings reported in the following sections, significant results can only identify associations. In presenting these findings we have nonetheless chosen to interpret some associations causally, while recognising that these interpretations must be treated with caution.

3. STUDENT VIEWS OF THE ROLE OF PROOF

Figure 1 shows the distribution of students' responses according to the roles students ascribed to proof, gleaned from their answers to the question *"what is proof for?"*. Each response was coded according to references to truth, explanation and discovery as described above (see Table 1). Any description which mentioned several proof roles received multiple codes. The distribution of responses is shown in Figure 1 below.

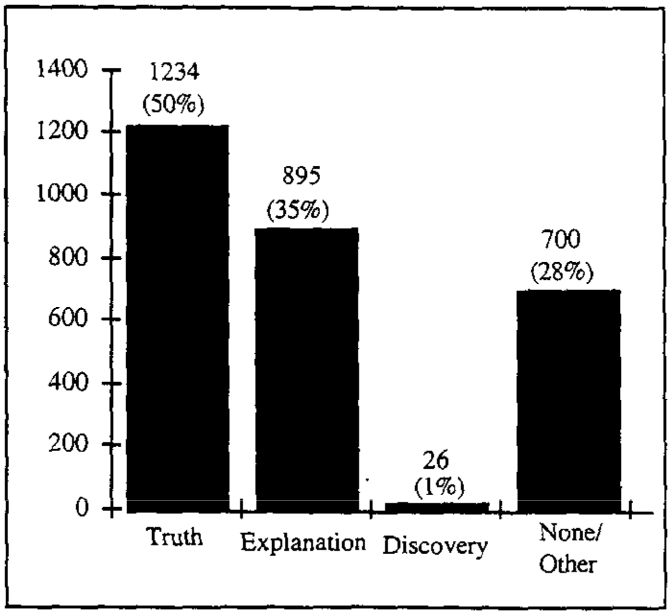


Figure 1: Distribution of students' responses to "what is proof for?".

Figure 1 shows that the most common response was that proof involved establishing the truth of a statement, a role mentioned by a little over 50% of all students. The function of communicating or explaining results, coded as explanation, was mentioned by 35% of students, while responses indicating some kind of discovery role for proof were very rare, referred to by only 1% of students. 28% of students either gave no answer or gave one that made little sense (coded as none/other), suggesting that a sizeable minority of these able students had no clear idea of what was meant by proof or what it was for.

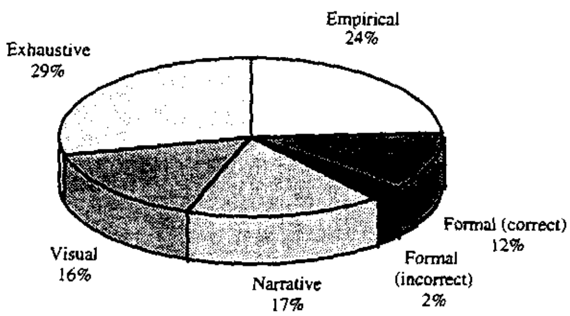
F1. Students are most likely to describe proof as about establishing the truth of a mathematical statement, although a substantial minority ascribe it an explanatory function and a further large number have little or no idea of the meaning of proof and what it is for.

4. DESCRIPTIVE STATISTICS: STUDENT RESPONSES TO THE MULTIPLE-CHOICE QUESTIONS

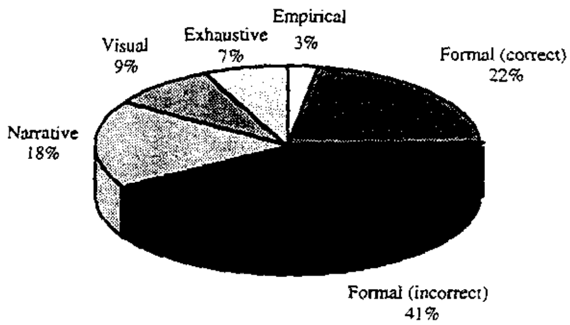
The survey included a total of 6 multiple-choice questions: for each of algebra and geometry, two concerning direct proofs and one the refutation of a false conjecture. In each question, students were asked to make two choices from the range of arguments presented: the argument closest to what they would do (designated below as choice for *own approach*); and the argument they believed would receive the *best mark*. All the frequency distributions are presented below in pie charts to facilitate comparisons.

DIRECT PROOFS: DISTRIBUTION OF CHOICES

First we focus on direct proofs. The following Figures 2 to 5 present the distributions of students' choices in response to the 4 multiple-choice questions concerning direct proofs.



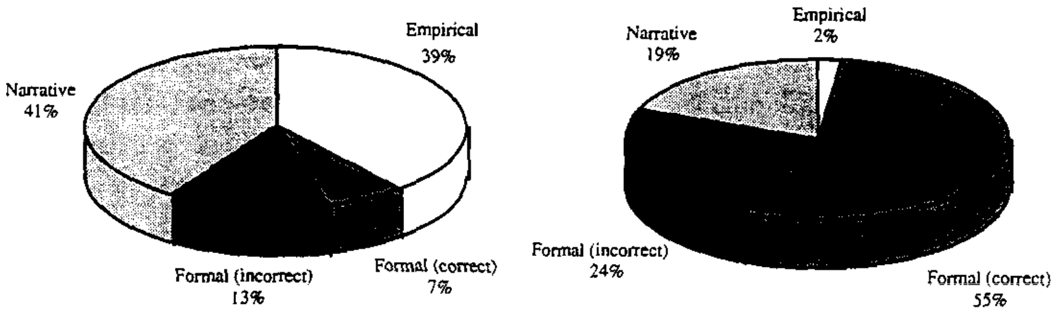
a) Own approach



b) Best mark

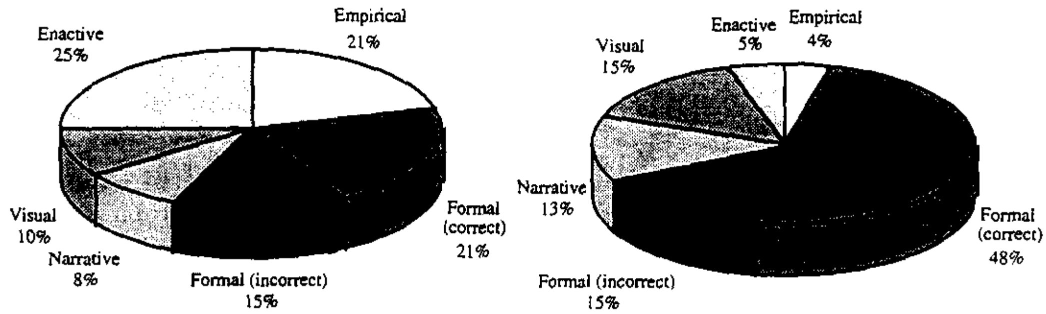
Figure 2: Distribution of students' choices in A1[13].

13 In the interest of brevity, the word 'analytical' is dropped from the description of the forms of proof, analytical formal, (correct and incorrect) and analytical narrative.



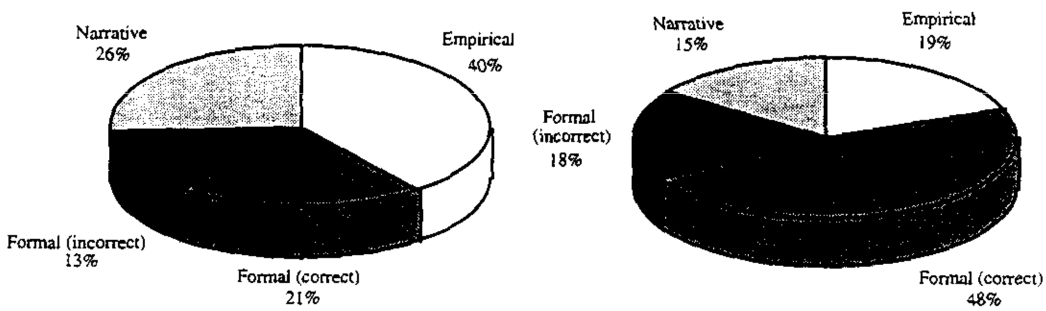
a) Own approach b) Best mark

Figure 3: Distribution of students' choices in A6.



a) Own approach b) Best mark

Figure 4: Distribution of students' choices in G1.



a) Own approach b) Best mark

Figure 5: Distribution of students' choices in G6.

The figures above all show marked variation between the choices made for own approach and for best mark, with differences that are highly significant for each question (A1: χ^2 = 1741.5, df = 5, p < 0.0001; A6: χ^2 = 1891.2, df = 3, p < 0.0001; G1: χ^2 = 922.88, df = 5, p < 0.0001; G6: χ^2 = 466.18, df = 3 p < 0.0001). This leads us to conclude:

F2. The argument selected as a proof of a conjecture is influenced by whether the choice is for the student's own approach or for the best mark.

Turning to the forms of argument that students considered would be awarded the *best mark*, we can report:

a) Formal presentation was always the most popular choice for the best mark, with the two formal forms of presentation (correct and incorrect) accounting for the following percentages of choices: A1: 63 %; A6: 79%; G1: 63%; and G6: 66%.

b) Students were much more likely to choose the formal option for best mark than for their own approach, a difference which is highly significant for all direct proofs. (A1: $\chi^2 = 763.73$, df = 1, $p < 0.0001$; A6: $\chi^2 = 827.5$, df = 1, $p < 0.0001$; G1: $\chi^2 = 187.82$, df = 1, $p < 0.0001$; G6: $\chi^2 = 218.76$, df = 1, $p < 0.0001$).

Taken together, we conclude:

F3. Students believe that a formal presentation of a proof will receive the best mark.

We also found that when choosing empirical arguments, students were much more likely to choose them as the argument most closely resembling their own approach than for the best mark, with differences which are highly significant: (A1: $\chi^2 = 419.92$, df = 1, $p < 0.0001$; A6: $\chi^2 = 777.33$, df = 1, $p < 0.0001$; G1: $\chi^2 = 273.77$, df = 1, $p < 0.0001$ (empirical); G1: $\chi^2 = 339.33$, df = 1, $p < 0.0001$ (enactive); G6: $\chi^2 = 162.24$, df = 1, $p < 0.0001$), we report

F4. Students are significantly more likely to select empirical arguments for their own approach than to receive the best mark.

Looking in more detail at the form of students' choices for their *own approach to algebra proofs* (Figures 2a, 3a), the two formal forms, i.e. symbolic, were the least popular (A1: 14%; A6: 20%). Students most frequently chose arguments presented in prose-form (A1: exhaustive and narrative forms combined account for 46% of choices; A6: the narrative form accounts for 41% of choices). We therefore report:

F5. In algebra proofs, the most popular choice of presentation for a student's own approach has a prose-form, while the least popular is symbolic.

Finally, considering students' choice for *best mark in algebra proofs* (Figures 2b, 3b), it is clear that the empirical form was the least frequently selected (A1: 3%; A6: 2%) indicating:

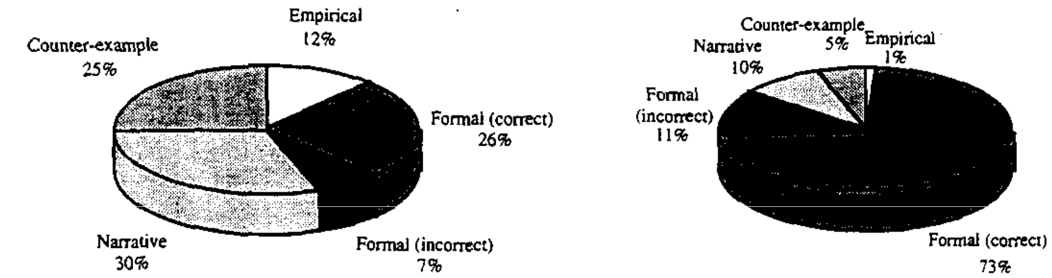
F6. An empirical verification of an algebra proof is very unlikely to be chosen to receive the best mark.

The findings for students' choices in geometry (Figures 4a, 6a) are less clear-cut than in algebra, although we note that a larger proportion of students chose the formal form as their own approach in this domain than in algebra (G1: formal accounts for 36% of student choices; G6: formal accounts for 34% of student choices). We therefore report:

F7. A formal presentation of a proof is a more popular choice for a student's own approach in geometry than in algebra.

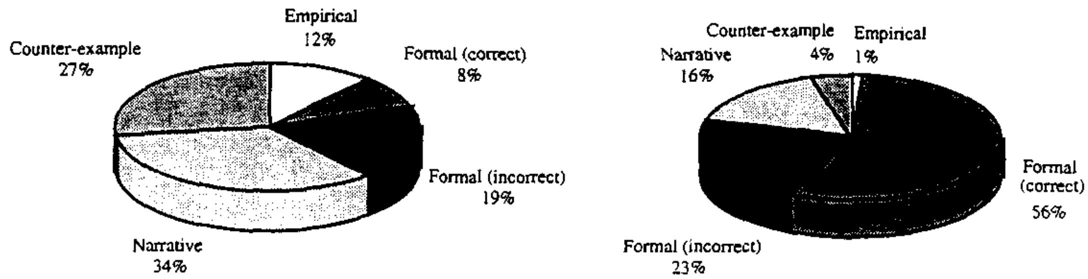
RESPONDING TO A FALSE CONJECTURE: DISTRIBUTION OF CHOICES

Next, we focus on the arguments selected by the students as their own approach and for best mark when faced with a *false conjecture,* (questions A5 and G5). The following Figures 6 and 7 show the distribution of student choices.



a) Own approach b) Best mark

Figure 6: Distribution of students' choices in A5.



a) Own approach b) Best mark

Figure 7: Distribution of students' choices in G5.

The pattern of choices again indicates significant variation between the argument chosen for own approach and for best mark (A5: $\chi^2 = 1351.5$, df = 4, $p < 0.0001$; G5: $\chi^2 = 1531$, df = 4, $p < 0.0001$), and finding F2 is generalised as below:

F2. The form of argument selected to refute a proof or conjecture is influenced by whether the choice is for the student's own approach or for the best mark.

We were particularly interested in students' evaluations of the role of a simple counter-example in refuting a false conjecture. In both domains, we found that students were very unlikely to choose a simple counter-example for best mark, although they were significantly more likely to do so for their own approach (algebra: $\chi^2 = 314.28$, df = 1 $p < 0.0001$; geometry: $\chi^2 = 382.54$, df = 1 $p < 0.0001$).

F8. Students are very unlikely to choose a simple counter-example for best mark, although they are significantly more likely to do so for their own approach.

We also noticed that, as with the direct proofs, formal presentation was the most popular choice for best mark. The two formal forms accounted for 84% of choices in A5 and 79% in G5, and were chosen significantly more for best mark than for a student's own approach (A5: $\chi^2 = 544.74$, df = 1 $p < 0.0001$; G5: $\chi^2 = 576.38$, df = 1, $p < 0.0001$). This provides further evidence for finding F3 which is extended as below:

F3. Students believe that proving or refuting a conjecture by means of a formally-presented analytic argument will receive the best mark.

We also found once again that students were significantly more likely to choose empirical arguments for their own approach and only rarely believed such arguments would receive the best mark: (A5: $\chi^2 = 204.59$, df = 1, $p < 0.0001$; G5: $\chi^2 = 181.73$, df = 1, $p < 0.0001$), confirming that finding F4 applies also when students are faced with false conjectures.

CHOOSING A PROOF WITH A CORRECT CONCLUSION

Finally we classified all the proofs presented in the multiple-choice questions in terms of whether or not they were *correct* from a mathematical perspective. Not surprisingly, when faced with false conjectures, most students selected an option which correctly refutes the conjecture for their own approach and for best mark (A5: 81%; G5: 69% for own approach, and A5: 88%; G5: 76% for best mark), with a slightly higher percentage choosing a correct argument for the best mark than their own approach. For direct proofs, students again were more likely to choose a correct proof than a false one for best mark (A1: 56%; A6: 74%; G1: 76%; G6: 63%) — the combined influence of their preference for the formal option and the unpopularity of empirical arguments for best mark. When choosing for own approach however, where empirical arguments were more popular, (marginally) less students selected correct proofs than incorrect ones; the one exception being question A1 (A1: 74%; A6: 52%; G1: 61%; G6: 53%).

F9. Students are more likely to choose a correct argument for best mark than for their own approach.

FACTORS ASSOCIATED WITH CHOICES: IDENTIFYING TRENDS FROM THE DESCRIPTIVE STATISTICS

We have reported that the distribution of choices of argument is significantly influenced by the criterion for the choice — own approach or best mark. However, these two distributions are not completely independent, as evidenced from the construction of cross-tables of student choices for each question. Significant correlations are found for both direct proofs and for refutations (A1: $\chi^2 = 437.5$, df = 25, $p < 0.0001$, $r = 0.3911$; A5: $\chi^2 = 1160.6$, df = 16, $p < 0.0001$, $r = 0.571$; A6: $\chi^2 = 193.0$, df = 9, $p < 0.0001$, $r = 0.276$; G1: $\chi^2 = 847.4$, df = 25, $p < 0.0001$, $r = 0.509$; G5: $\chi^2 = 826.8$, df = 16, $p < 0.0001$, $r = 0.514$; G6: $\chi^2 = 262.8$, df = 9, $p < 0.0001$, $r = 0.323$).

F10. The argument believed to receive the best mark influences a student's choice for his/her own approach.

We also investigated whether the validity rating, VR, accorded to any argument was associated with the students' choice for their own approach. For every proof in A1, A6, G1, G6, we compared the distribution of VR's, (0,1,2), for students who chose this proof as their own approach with those who did not. The pattern proved to be completely consistent for every question: if the argument was correct, students who chose it obtained a higher VR than those who made other choices; if the argument was incorrect, the reverse was the case. Rather than produce the evidence for each question, we illustrate this finding with two examples.

For A1, 73% of those who chose the formal correct argument as their own approach scored a VR for that proof of 2; that is, they correctly rated its validity. By comparison, only 36% of the students who had chosen this argument scored its validity correctly. This and other comparisons for correct arguments are illustrated below:

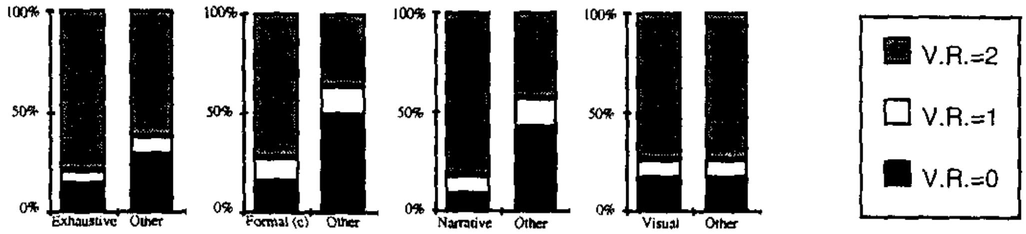


Figure 8(a) Validity Ratings for correct arguments in A1.

By contrast, only 37% of those who chose the (incorrect) empirical argument for A1 gave it a completely correct validity rating, as compared to 65% of the rest, who clearly recognised its limitations. This comparison and the percentage responses for the incorrect formal option are illustrated below:

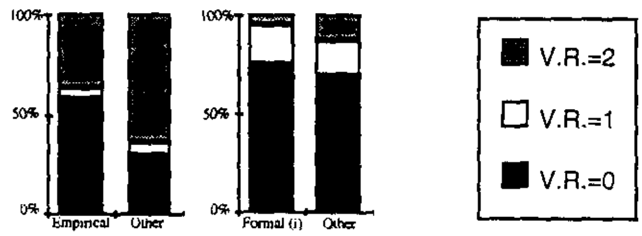


Figure 8(b) Validity Ratings for incorrect arguments in A1.

Although indications of the effect of an appreciation of the generality of an argument are discernible in answers to every question, the strength of its influence is diminished by the difficulty of the question, particularly in algebra. For example, in A6, 56% of students who chose an empirical argument actually assessed its validity quite correctly, i.e. they knew it was not general but nonetheless admitted that this approach would be closest to what they would have to do.

Overall, we report the following:

F11. An argument whose generality is correctly appreciated is more likely to be chosen as a student's own approach than one that is not.

Students also rated every argument in terms of how far it convinced them and explained why the conjecture was true. We investigated whether this explanatory power, EP, of a proof was associated with a student's choice for his/her own approach. Again, the pattern was completely consistent: for every question and every proof (whether correct or incorrect), students who chose a proof as their own approach gave it a higher EP score than those who did not. This was particularly marked for the assessments of narrative arguments. 59% of the students who chose this form in A1 assigned it a maximum EP of 2, as compared to only 11% of those who did not choose it. While only 8% of the students who chose it gave it an EP of 0, in contrast to 33% who had made a different choice. Similarly for A6, 60% of those who chose the narrative form gave it the maximum EP, compared to only 24% amongst those who did not choose it. Rather than produce these statistics for each question, we simply illustrate this consistent trend for A1 and A6.

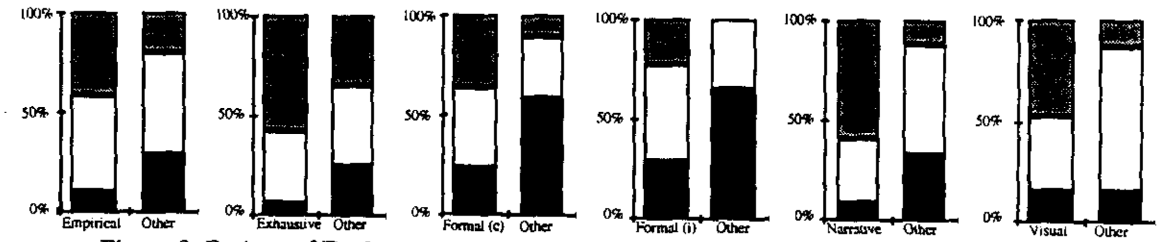


Figure 9: Ratings of Explanatory Power of arguments in A1.

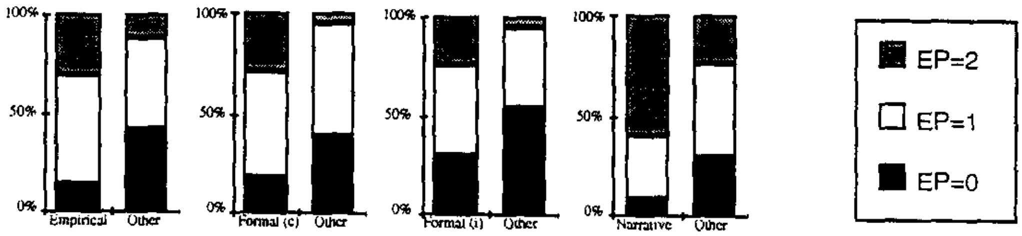


Figure 10: Ratings of Explanatory Power of arguments in A6.

So we report:

F12. An argument felt to convince or explain is more likely to be selected as a student's own approach than one that is not.

5. MULTILEVEL MODELLING: STUDENT RESPONSES TO THE MULTIPLE-CHOICE QUESTIONS

Our results indicate that, in all the multiple-choice questions, student choices for their own approaches correlate with their choice for best mark, and their views about the generality and explanatory role of the arguments. However, simply looking at whether one input variable correlates with an outcome variable may give a misleading picture, as interactions between input variables will not be identified and any clustering of responses associated with shared classroom experiences cannot be taken into account. To explore systematically which factors are associated with differences in performance and whether student output measures vary from school to school, multilevel modelling techniques were used to take into account the two-level structure of the data set: with school (including school, curriculum, teaching and teacher factors) at Level 2[14] (see Table 5) and students at Level 1.

14 Since we obtained responses from 2 classes in only 4 schools, it is impossible to distinguish between school and class effects.

A multinomial model of students' choices for their own approach[15] was constructed for each of the six multiple-choice questions, so that variables significantly associated with student choice patterns could be isolated and their effects in relation to students' preferences identified. The construction of a multinomial model of a categorical outcome involves selecting one category as a fixed base or *comparison* category and comparing responses to this with responses to the other categories[16]. In all cases we chose the empirical category as the basis for comparison, except in G5, where this category did not exist and the naive category was used. Details of the models applied can be found in Appendix 5.

MODELS OF STUDENT CHOICES IN DIRECT PROOFS

Table 6 below lists the χ^2 values of all the variables significantly associated with student choices for their own approach to proving the statements in A1, A6, G1 and G6.

	Question			
	A1	A6	G1	G6
Variables				
Level 1				
Views of role of proof				
Truth	69.88** (5)			12.20* (3)
Discovery	17.44* (5)		117.7*** (5)	
Explanation	16.46* (5)			14.03* (3)
Student characteristics				
Sex	30.99** (5)		21.44* (5)	8.74* (3)
KS3 score	313.5*** (5)	119.4*** (3)	274.7*** (5)	57.38** (3)
Responses to questionnaire				
Best mark	294.6*** (5)	37.3** (3)	259.1*** (5)	80.27***(3)
Validity rating	189.8*** (5)	126.0*** (3)	181.1*** (5)	153.2*** (3)
Explanatory power	823.2*** (5)	379.6*** (3)	793.0*** (5)	641.3*** (3)
Level 2				
School factors				
Selection procedures	25.73* (5)		20.5* (5)	
Curriculum factors				
Examination syllabus	52.96** (15)			
Main textbook/scheme	45.56**	24.1** (12)		
Hours of mathematics	27.05* (5)	27.2** (3)		
Approaches to teaching proof				
Write algebra proofs		11.8* (3)		

Notes:
a. * = p<0.01; ** = p<0.001; ***= p<0.0001.
b. df shown in brackets.

Table 6: χ^2 values showing the effects of significant variables underlying choices.

Table 6 indicates that only four variables (shown in bold) are associated with student choices in all the 4 questions, and all of these are student factors. The first is a student's Key Stage 3 test score, which is not altogether surprising as this provides a general input measure of mathematical attainment.

F13. Key Stage 3 test score influences the choice of argument for a student's own approach.

15 Models were not constructed for students choices for the best mark.

16 In fact, estimates of the logarithms of the ratio of the number of students choosing any category to the number of students choosing the comparison category are obtained.

The other three variables are more interesting as they involve student responses to other questionnaire items — their choice for best mark, and their assessments of the scope of validity and explanatory power of their choice. These correlations confirm the trends picked out in findings F10, F11 and F12 from the descriptive statistics, and so point to the highly significant nature of these factors — they remain important, even when other variables are included in the model.

There are however other student factors which influence choices in the majority of questions. Girls and boys had significantly different choice patterns in all questions, except the more unfamiliar algebra proof.

F14. In most cases, girls and boys choose different arguments as their own approach to prove a statement.

Student views of the role of proof also influence choices in three out of four of the multiple-choice questions.

F15. Student views of the role of proof influence their choice of argument for their own approach, except when presented with an unfamiliar algebra proof.

Turning to Level 2 variables, it is evident from Table 6 that these are rarely, if ever, associated with student choices in geometry and, for the unfamiliar geometry question, no Level 2 variables prove to be significant. By contrast, in algebra, curriculum variables are associated with choices, with the main textbook and the hours of mathematics teaching per week significant in both questions.

F16. Curriculum factors influence the proofs chosen in algebra, with the main textbook and the hours of mathematics teaching each week exhibiting the most consistent effects.

F17. Choices of proof in geometry are predominantly associated with student rather than school, curriculum and teacher factors.

Table 6 above shows the variables associated with student choices, but not whether their effect extends to all, or only some, of the arguments. Table 7 below presents the estimated effects of the significant variables on each category of argument in algebra, and Table 8 shows the same information for geometry.

		Forms of proof				
		Exhaustive	Formal (c)	Formal (i)	Narrative	Visual
Base group ratios	A1	-1.73 (0.23)	-2.92 (0.35)	-5.03 (0.67)	-3.86 (0.36)	-2.18 (0.29)
	A6	—	-4.35 (0.32)	-2.40 (0.20)	-2.27 (0.17)	—
Variables						
Level 1						
Views of role of proof						
Truth	A1	0.27 (0.10)	0.27 (0.13)	0.70 (0.30)		0.29 (0.12)
Discovery	A1				0.82 (0.41)	
Student characteristics						
Sex	A1		-0.36 (0.12)			
KS3 score	A1	0.39 (0.07)	0.57 (0.09)		0.53 (0.08)	0.30 (0.09)
	A6	—	0.29 (0.12)	0.58 (0.11)	0.46 (0.08)	—
Responses to questionnaire						
Best mark	A1			1.41 (0.36)	0.37 (0.10)	1.94 (0.17)
	A6	—	1.69 (0.26)	0.36 (0.17)	0.39 (0.16)	—
VR score	A1	0.21 (0.06)	0.59 (0.08)	0.58 (0.23)	0.79 (0.08)	
	A6	—	0.43 (0.11)		0.61 (0.06)	—
EP score	A1	0.62 (0.13)	1.03 (0.08)	1.52 (0.21)	1.25 (0.08)	0.93 (0.09)
	A6	—	1.07 (0.16)	1.01 (0.10)	0.92 (0.07)	—
Level 2						
School factors						
School selection	A1		0.69 (0.26)	1.06 (0.32)	0.64 (0.24)	
Curriculum factors						
Hours of mathematics	A1	0.39 (0.12)	0.89 (0.19)		0.41 (0.19)	0.32 (0.16)
	A6	—	0.39 (0.17)	0.60 (0.15)	0.35 (0.12)	—
Examination syllabus						
SEG	A1	0.40 (0.19)				0.66 (0.23)
London	A1	0.49 (0.20)		-0.76 (0.35)		
Main textbook/scheme						
Holderness	A1					-0.61 (0.24)
Vickers	A1					-0.54 (0.24)
Approaches to teaching proof						
Write algebra proof	A6	—		0.46 (0.17)		—

Notes:

a. The comparison category for both questions is the empirical form.

b. Visual and exhaustive forms in A1 only.

c. Standard errors in brackets.

d. Some variables may improve the model overall but significant estimates for particular categories may not be obtained. These variables are not shown here.

e. — indicates proof form not available for this question.

Table 7: The estimated effects of the significant variables on student choices in the algebra multiple-choice questions (direct proofs).

		Forms of proof				
		Enactive	Formal (c)	Formal (i)	Narrative	Visual
Base group ratios	G1	-0.84 (0.15)	-2.90 (0.19)	-1.87 (0.17)	-3.47 (0.25)	-2.54 (0.23)
	G6	—	-3.35 (0.23)	-2.68 (0.20)	-2.14 (0.17)	—
Variables						
Level 1						
Views of role of proof						
Truth	G6	—	0.33 (0.12)			—
Explanation	G6	—	0.35 (0.11)			—
Student characteristics						
Sex	G1	0.44 (0.11)				
	G6	—	0.60 (0.20)			—
KS3 score	G1	0.41 (0.08)	0.57 (0.08)	0.37 (0.09)	0.54 (0.10)	
	G6	—	0.33 (0.09)	0.35 (0.19)	0.29 (0.08)	—
Responses to questionnaire						
Best mark	G1	1.28 (0.24)	0.95 (0.15)	0.49 (0.19)	1.12 (0.19)	1.86 (0.21)
	G6	—	1.12 (0.17)	0.98 (0.20)	0.47 (0.17)	—
VR score	G1	-0.33 (0.05)	0.50 (0.06)	1.88 (0.58)	0.58 (0.09)	
	G6	—	0.38 (0.07)		0.28 (0.06)	—
EP score	G1	0.53 (0.07)	1.14 (0.08)	1.16 (0.08)	1.05 (0.12)	1.01 (0.10)
	G6	—	1.10 (0.08)	1.23 (0.09)	1.00 (0.08)	—
Level 2						
School factors						
School selection	G1	-0.38 (0.18)				

Notes:

a. The comparison category for both questions is the empirical form.
b. Visual and enactive forms in G1 only.
c. Standard errors in brackets.
d. Some variables may improve the model overall but significant estimates for particular categories may not be obtained. These variables are not shown here.
e. — indicates proof form not available for this question.

Table 8: Estimated effects of the significant variables on student choices in the geometry multiple-choice questions (direct proofs).

To explain how to interpret these estimates, we describe the model in detail for the familiar algebra question, A1. For this question, we identified that six Level 1 variables and four Level 2 variables are associated with student choices of argument: views of the role of proof; student sex; Key Stage 3 test score; best mark; validity rating (VR); explanatory power (EP); school selection procedures; examination syllabus; main textbook/scheme; and hours of mathematics teaching per week. Particular values of these 10 variables were therefore chosen to define a base group of male students with an average Key Stage 3 score of 6, coming from non-selective schools, not following one of the popular examination syllabi (SEG, MEG or London) or using a popular mathematics textbook (SMP, Holderness, Vickers or Rayner) and receiving the average hours of mathematics teaching (3 hours) per week. This group was further defined by the following responses to other questionnaire items: they offered no view of the role of proof, chose different options for their own approach and best mark and scored 0 for the VR and the EP of their choice of argument for their own approach. The estimates for this base group, *the base group ratios,*[17] were all negative (exhaustive/empirical, -1.73; formal correct/empirical -2.92; formal incorrect/empirical -5.82; narrative/empirical -3.86; visual/empirical -2.18). This suggests that the empirical option was likely to be the most popular choice for these students, which stands in contrast to the finding that overall the exhaustive option was the most frequently selected.

17 In fact, the logarithm of the ratio of the number of these students choosing a particular proof form to the number choosing the empirical.

For the explanatory variables, a positive estimate indicates an increase in the likelihood of choosing a particular category in preference to the comparison category, while a negative estimate indicates a decrease[18]. So, for example, by considering the estimates associated with the formal correct option in A1, we find that the students most likely to choose it were those who: believed proof to have a truth role (increases the base group ratio by 0.27), were male (being female decreases the base group ratio by 0.36), had a Key Stage 3 test score of 8 (increases the base group ratio by 2 * 0.57), had a VR and EP score of 2 for this proof (increases the base group ratio by 2 * 0.59 and 2 * 1.03 respectively), and attended a selective school (increases the base group ratio by 0.69). The estimate for students defined in these ways is 4.98, suggesting that students in this group are far more likely to choose the formal correct than the empirical option in A1.

A similar analysis was undertaken for each multiple-choice question leading to the group of findings reported below, beginning with those where the variables have similar effects in both algebra and geometry.

Tables 7 and 8 show that a student's Key Stage 3 score has a significant and positive effect for nearly all proof forms in all the four questions, excluding only the formal incorrect in A1 and the visual in G1. We can extend finding F13 to the following:

F13. Key Stage 3 test score influences the choice of argument for a student's own approach; as this score increases so does the student's preference for an argument which is not empirical.

Positive estimates for the variable, best mark, are also obtained for all proof forms in all the four questions (except for the exhaustive and the formal correct options in the familiar algebra question). This general pattern provides evidence of the direction of influence of this variable, hence finding F10 can be restated as:

F10. Students are more likely to choose an argument for their own approach if they believe it will receive the best mark.

In the majority of cases, positive estimates for the variable, VR score, are obtained for all proof forms in the four questions, indicating that a student's preference for a particular argument increases when its generality is assessed correctly[19]. A negative estimate is obtained for the incorrect enactive argument in G1, indicating that students were less likely to choose this option when they evaluated its scope of validity correctly — that is, when they knew it was not general. No significant estimates are obtained for the visual option, where the scope of validity is rather ambiguous: it is possible that students choosing the visual option were attracted to it for reasons other than their assessment of

18 For a categorical variable like student sex, the estimate represents the ratio for one category divided by the ratio for the other (so the predicted effect for girls is the logarithm of the ratio for girls divided by the ratio for boys). For a continuous variable, the estimate represents the predicted change in choice preference for a one unit change in the continuous variable (e.g., a one level increase in Key Stage 3 score).

19 Oddly, we find that on the questions with familiar mathematical content in both algebra and geometry, this finding also holds for the formal incorrect proof.

its generality. This adds support to finding F11, with the proviso that it does not hold for visual or incorrect arguments.

Positive estimates for the variable, EP are associated with every proof form for all four questions. This adds support to finding F12, which we rewrite below:

F12. An argument felt to convince or explain is more likely to be selected as a student's own approach than one that is not, and the likelihood increases still further if it does both.

Where significant effects for the variable role of proof are found, they are always positive. This suggests that an awareness of *any* role for proof decreased any student propensity towards empirical arguments, although this effect does not reach a level of significance for all roles and for all choices. We therefore replace F15 by the following:

F15. In most cases, student views of the role of proof influence their choice of argument for own approach and, in particular, students who have some idea of the role of proof are less likely to choose empirical arguments than those who do not.

The variable, student sex, is associated with different choices in three of the four questions, but for each question significant estimates are observed with respect to one choice only and patterns are not consistent across questions. The negative effect observed in the familiar algebra question shows girls were less likely than boys to choose the formal correct option, while in the unfamiliar geometry question a positive estimate is obtained for the formal correct argument. We would have therefore to conclude the specific argument favoured by girls and boys differed according to the mathematical content of the question presented.

Level 2 variables, especially curriculum variables, account for variation in the choices in algebra, but rarely in geometry. Even in algebra, only one variable, hours of mathematics teaching per week, has a consistent effect across both questions with significant and positive estimates for almost all choices.[20] We therefore rewrite F16 as follows:

F16. Curriculum factors influence the arguments chosen for algebra proofs, with the main textbook and the hours of mathematics teaching each week exhibiting the most consistent effects. In particular, increasing the number of hours of mathematics teaching each week reduces the likelihood of students choosing an empirical argument.

The general inconsistency of the effects of curriculum variables across questions suggests that choices result less from differences in emphasis on the process of proving and more from the coverage of the particular mathematical content of the question.

20 Other findings are more specific to particular questions and forms of proof: e.g. students following the Holderness or Vickers textbooks were less likely to opt for a visually-presented algebra proof than those using an alternative study scheme.

SCHOOL DIFFERENCES

We now turn to look at variation in response between students from different schools. We were interested to explore whether by including all the significant input variables in a model, we would no longer find any differences in the choice patterns of students in different schools — or, put another way, whether even after adjustment, schools varied around their predicted ratios. This would suggest that schools were influencing students' choices in ways not captured by the explanatory factors.

In Table 9 below, the estimates of the amounts by which schools varied around their predicted ratios for each argument in algebra are shown in bold along the leading diagonal. The covariances between arguments are also presented in order to assess the association between preferences within schools. For example the positive correlation, 0.99, between the exhaustive and formal correct choices in question A1 shows that in schools where students were more likely to favour the exhaustive over the empirical, they were also likely to favour the formal correct.

Forms of proof		Exhaustive	Formal (c)	Formal (i)	Narrative	Visual
Exhaustive	A1	**0.05(.04) NS**				
	A6	—				
Formal (c)	A1	0.14(.03) *.99*	**0.40(.09)**			
	A6	—	**0**			
Formal (i)	A1	n/a	n/a	**n/a**		
	A6	—	0	**0.13(.07) NS**		
Narrative	A1	0.14(.04) *.90*	0.46(.09) *.99*	n/a	**0.52(.010)**	
	A6	—	0	0.02(.05) *.21*	**0.09(.04) NS**	
Visual	A1	0.01(.03) *.08*	0.13(.06) *.75*	n/a	0.13(.06) *.63*	**0.08(.06)NS**
	A6	—	—	—	—	—

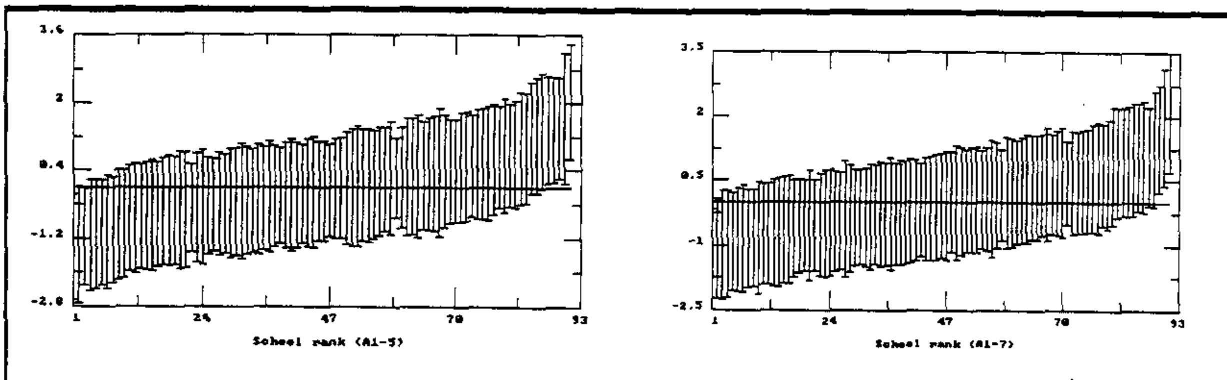
Notes.
a. n/a indicates too few responses to obtain an estimate.
b. Exhaustive and Visual forms in A1 only.
c. Correlation coefficients shown in italics.
d. NS indicates variation is not significant.
e. — indicates proof form not available for this question.

Table 9: Random effects (variance and covariance estimates) at school level in the algebra multiple-choice questions.

The estimates of variance for the familiar algebra question A1 presented in Table 9 (on the leading diagonal in bold) indicate that students' preferences for some arguments did indeed differ according to which school they attended: that is, there remained some variation in student choices that could not be accounted for by the input variables. However, this school variation reached a level of significance only with respect to two choices: preferences for narrative over empirical (0.52) and formal correct over empirical (0.40). Figure 11 illustrates the 95% uncertainty intervals around the residual estimates for each class[21] for these two comparisons. Both plots show considerable overlap between the schools. However, the gradient of the estimates appears to rise more sharply at the extremes of each plot, especially at the upper extreme — the schools with large positive ratios. In this handful of schools, the ratio of students choosing a formal correct or narrative argument, both analytical forms of proofs, in preference to the empirical was better than would be expected. To investigate further the reasons for these between-school differences, we selected from each plot the five schools with the largest positive

21 The extent to which the actual log ratios differed from the log ratio predicted by the fixed part of the models.

estimates and the five with the largest negative estimates to be added to a sample from which a number of schools were chosen for case study in the second phase of our research.



Formal correct/Empirical in A1 **Narrative/Empirical in A1**

Figure 11: School deviations from predicted ratios: 95% uncertainty intervals around estimates for each school.

If we look at the covariance estimates presented in Table 9, we find strong associations between the school predictions for the exhaustive, formal correct and narrative arguments (0.99 exhaustive/formal, 0.90 exhaustive/narrative and 0.99 formal/narrative). This implies that in schools where the ratio of exhaustive to empirical choices was greater than predicted, the proportions choosing the formal correct and narrative was also larger than expected. Moderate associations are found between the visual and the formal correct arguments (0.75) and the visual and narrative (0.63). There is no significant association between the exhaustive and the visual forms.

For the unfamiliar question, A6, there was no significant variation at the school level after adjustment for the significant variables, with zero estimates obtained for the formal correct choice and very small estimates with large standard errors for the others. This suggests that, after taking into account differences between students, schools and curriculum, choices did not differ significantly according to the school attended.

F18. Although there is little between-school variation in the arguments selected as proofs of a conjecture, there are some schools where the students' preferences for analytical arguments are greater than predicted.

Moving on to the domain of geometry (see Table 10 below) and looking first at the familiar question, G1, the estimates of the random effects indicate that schools vary significantly around their predicted ratios for all proofs except the narrative.

Forms of proof		Enactive	Formal (c)	Formal (i)	Narrative	Visual
Enactive	G1	**0.15 (.06)**				
	G6	—				
Formal(c)	G1	-0.05 (.04) *-.35*	**0.12 (0.05)**			
	G6	—	**0.89 (0.18)**			
Formal(i)	G1	-0.05 (.05) *-.31*	0.13 (0.05) *.88*	**0.19 (0.08)**		
	G6	—	0.58 (0.12) *.98*	**0.39 (0.12)**		
Narrative	G1	0.001 (.04) *.03*	-0.02 (.04) *-.75*	-0.004 (.05) *-.11*	**0.01 (0.05) NS**	
	G6	—	0.34 (.09) *.65*	0.21 (0.07) *.60*	**0.31 (0.08)**	
Visual	G1	-0.03 (.06) *-.17*	-0.07 (.05) *-.41*	-0.14 (.07) *-.64*	0.08 (0.06) NS	**0.23 (.10)**
	G6	—	—	—	—	—

Notes.
a. Enactive and Visual forms in G1 only.
b. Correlation coefficients shown in italics.
c. NS indicates variation is not significant.
d. — indicates proof form not available for this question.

Table 10: Random effects (variance and covariance estimates) at school level in the geometry multiple-choice questions.

Figure 12 below presents 95% uncertainty intervals around the residuals for each class for the remaining comparisons, and as in the corresponding algebra question, these plots clearly illustrate the extent to which schools overlap. Again, for us, the most interesting schools are located at either extreme of the horizontal axis since in these schools the pattern of choices differs most from that predicted. We noted the top and bottom five schools from each graph and added them to the sample from which we later selected schools for case study.

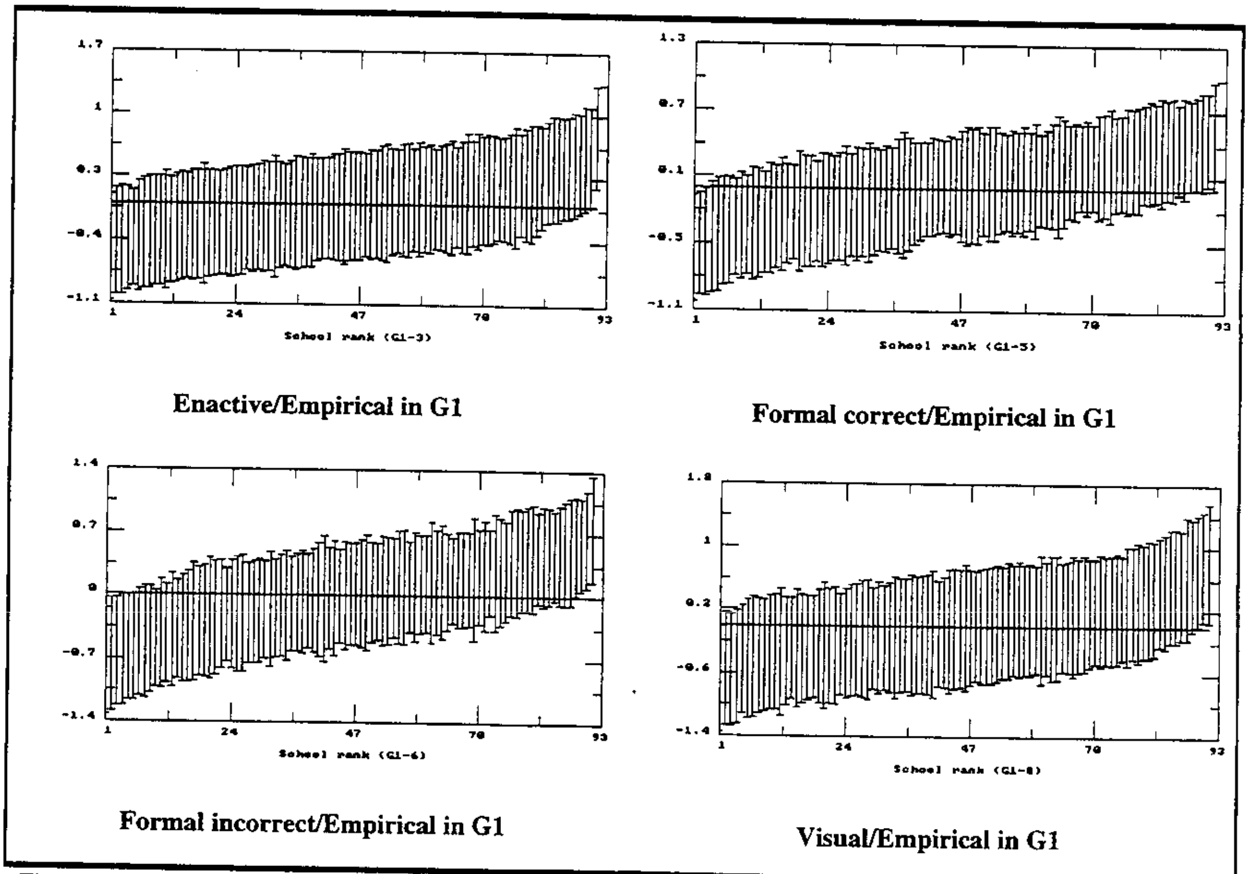


Figure 12: School deviations from predicted ratios: 95% uncertainty intervals around estimates for each school.

For the unfamiliar geometry question, G6, estimates of school-level variation are rather higher than observed for the other questions, especially with respect to the formal correct argument (0.89), although schools also varied around the predictions for formal incorrect (0.39) and narrative (0.31). This suggests that schools were varying in ways not modelled by the input variables. The 95% uncertainty intervals around the residuals for each class are presented in Figure 13 and illustrate clearly the differing performance in different classes. We therefore conclude:

F19. Student's choice for own approach varies more in geometry than in algebra according to school attended.

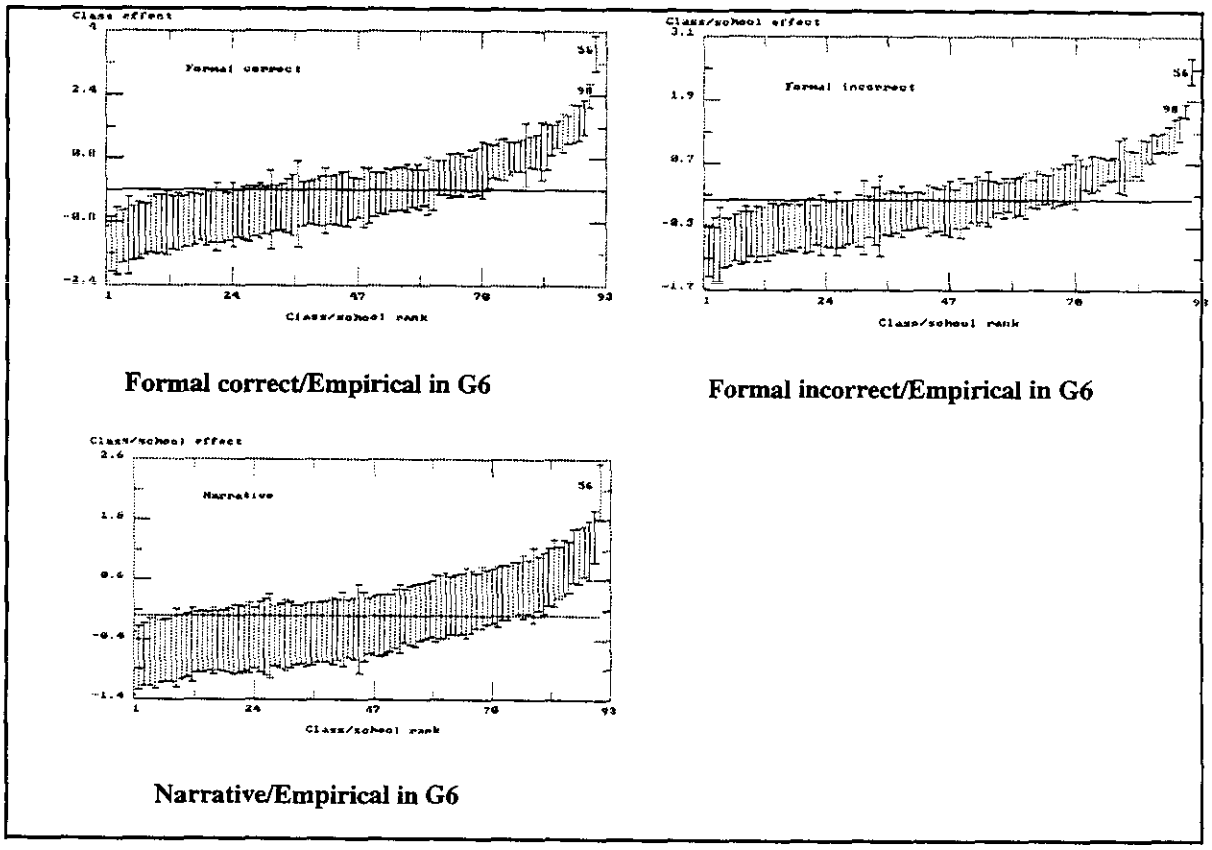


Figure 13: School deviations from predicted ratios: 95% uncertainty intervals around estimates for each school.

Turning to the covariance estimates, the strongest associations for G1 are between the two formal arguments, one correct and one incorrect (0.88), suggesting if one was popular in a school, so was the other. Weaker negative associations are found between the enactive and both formal arguments (-0.55 and -0.51) suggesting that in classes where students preferred the enactive argument to the empirical, neither formal option was popular. A negative association is also found between the formal correct and the narrative option (-0.75), suggesting that students in classes showing a propensity towards choosing a correct formal geometry proof were less frequently drawn to the narrative argument.

In terms of the covariation estimates for G6, we found positive associations between all 3 forms, and, as for G1, the strongest correlation between the two formal arguments (0.98).

Similar patterns are identifiable for the narrative and both formal forms, although the associations are more moderate (0.65 narrative/formal correct, 0.61 narrative/formal incorrect). Thus, in schools where fewer students selected the empirical, the popularity of all three arguments increased. Out of all these findings, we note:

F20. In geometry, the school attended can enhance students' preferences for a formal argument, regardless of whether or not it is correct.

MODELS OF STUDENT RESPONSES TO A FALSE CONJECTURE

Table 11 below lists the χ^2 values of all variables significantly associated with student responses to false conjectures. The comparison category used in these models was the empirical for the algebra question, A5, and the naive for the corresponding geometry question, G5, both of which are incorrect. The overall percentages of students selecting the comparison category was much smaller than in the case of the direct proof questions[22], which probably accounts for the greater number of significant variables isolated in the models below. Furthermore, two significant variables in the models for direct proofs (validation rating and explanatory power) were not available for these questions.

	Questions	
Variables	A5	G5
Level 1		
Views of role of proof		
Truth	750.3*** (4)	41.97***(4)
Social		106.7*** (4)
Student characteristics		
Sex	14.77** (4)	70.50*** (4)
KS3 score	2577.6*** (4)	1114.2*** (4)
Responses to questionnaire		
Best mark	5073.4*** (4)	675.0*** (4)
Level 2		
School factors		
School location	176.32*** (4)	342.1*** (4)
School sex	157.05*** (4)	52.35*** (4)
School selection procedures	147.32*** (4)	342.1*** (4)
Curriculum factors		
Examination syllabus	897.0*** (12)	180.5*** (12)
Main textbook/scheme	138.5***(16)	580.1*** (16)
Hours of mathematics		10.23* (4)
Approaches to teaching proof		
Through investigations	117.7*** (4)	
As separate topic	70.32*** (4)	148.6*** (4)
Read algebra/geometry proofs	458.7*** (4)	
Write algebra/geometry proofs	115.5*** (4)	15.37** (4)
Teachers' views of National Curriculum		
Emphasis on mathematical justification	175.2*** (8)	82.91*** (8)
Emphasis on formal proof	22.15** (8)	
Notes * = p<0.01; ** = p<0.001; ***= p<0.0001. df shown in brackets.		

Table 11: χ^2 values showing the effects of significant variables underlying choices.

22 For the direct proofs, all arguments led to a correct conclusion, but for false statements incorrect conclusions were drawn from the empirical and the formal incorrect arguments in A5 and for the naive and informal incorrect arguments in G5.

Table 11 shows that, in contrast to the situation for direct proofs, the significant variables underlying student responses to a false conjecture are largely the same in algebra and geometry. School variables feature more as significant predictors and include school location, sex of student intake and selection procedures. A number of factors related to the curriculum are also associated with choices in these two questions and include: the examination syllabus and main textbook used by the mathematics department, the approaches to teaching proof and teachers' views of coverage of mathematical justification and formal proof in the National Curriculum. The student factors found to be significant correspond more closely with the findings for direct proofs and again, include student views of the roles of proof, the sex of the student, their Key Stage 3 test score and their choice for the best mark.

F21. Significant influences on students' responses to a false conjecture include student, school, curriculum and teaching variables.

Table 12 below presents the estimated effects of the significant variables on each category of argument.[23]

23 Note that the standard errors for the base group ratios are very large for both questions. This is not surprising given the small number of students to which these estimates apply, but suggests that the base group ratios are rather unstable.

		Forms of proof			
		Formal (c)	Formal (i)	Narrative	Counter-example
Base group ratios	A5	-0.49 (0.47)	-2.11 (0.89)	0.18 (0.41)	n/a
	G5	-1.50 (0.72)	-1.00 (0.69)	-0.43 (0.50)	0.04 (0.47)
Variables					
Level 1					
Views of role of proof					
Truth	A5	0.56 (0.09)	0.84 (0.17)	0.52 (0.09)	0.70 (0.10)
	G5			0.22 (0.09)	
Social	G5			-0.29 (0.09)	-0.29 (0.10)
Student characteristics					
Sex	A5			-0.19 (0.09)	
	G5	-0.43 (0.12)	0.64 (0.13)	-0.20 (0.10)	0.38 (0.11)
KS3 score	A5	1.09 (0.07)	0.61 (0.13)	0.70 (0.07)	0.79 (0.08)
	G5	0.55 (0.08)	0.22 (0.09)	0.90 (0.07)	0.40 (0.08)
Responses to questionnaire					
Best mark	A5	0.41 (0.02)	0.27 (0.03)	0.13 (0.03)	0.54 (0.02)
	G5	2.39 (0.25)		1.83 (0.08)	1.44 (0.12)
Level 2					
School factors					
School location	A5	0.44 (0.11)	0.39 (0.20	0.37 (0.11)	
	G5	0.47 (0.14)	0.42 (0.17)	0.31 (0.12)	
School sex					
Girl only	A5				2.76 (0.28)
	G5	-0.62 (0.31)			0.62 (0.31)
Boy only	A5				0.59 (0.21)
	G5	-0.71 (0.21)			
School selection procedures	A5	0.52 (0.14)		0.48 (0.15)	
	G5	0.94 (0.17)	0.30 (0.23)	0.78 (0.15)	0.52 (0.16)
Curriculum factors					
Examination syllabus					
SEG	A5	0.43 (0.16)			-0.39 (0.16)
	G5				
MEG	A5		-0.72 (0.31)		-1.60 (0.19)
	G5			0.44 (0.19)	
London	A5	-0.91 (0.18)	-0.67 (0.29)	-0.51 (0.17)	-0.79 (0.17)
	G5				
Main textbook/scheme					
SMP	A5				0.94 (0.16)
	G5	0.35 (0.17)		0.54 (0.14)	0.42 (0.15)
Holderness	A5				
	G5	1.23 (0.17)		1.00 (0.16)	0.75 (0.17)
Vickers	A5			0.48 (0.15)	0.63 (0.19)
	G5				
Approaches to teaching proof					
Through investigations	A5	0.33 (0.12)		0.31 (0.13)	-0.32 (0.13)
	G5	-0.51 (0.14)		-0.45 (0.13)	
As separate topic	A5	0.62 (0.13)			
	G5				
Read algebra proofs	A5	0.54 (0.12)		0.70 (0.12)	0.96 (0.13)
	G5				
Write algebra proof	A5				-1.13 (0.13)
	G5				
Teachers' views of National Curriculum					
Emphasis on mathematical justification	A5			-0.64 (0.27)	
Emphasis on formal proof	G5	-1.02 (0.41)			

Notes:

a. The comparison category for A5 is the empirical form and for G5 the naive form.
b. Standard errors in brackets.
c. Some variables may improve the model overall but significant estimates for particular categories may not be obtained. These variable are not shown in this table.
d. n/a indicates the number of students in the base group choosing this option was too small to obtain an estimate.

Table 12: The estimated effects of the significant variables on student responses to false conjectures.

Considering first the Level 1 variables, Table 12 shows that a student's Key Stage 3 score has a significant and positive effect for all proof forms for both questions, indicating that:

F22. As students' Key Stage 3 test score increases, so do their preferences for correct refutations, but also their preferences for incorrect formally-presented arguments.

Positive estimates for the variable, best mark, are obtained for all arguments, except the formal incorrect in G5, which indicates finding F11 also applies to students' responses to false conjectures.

For the variables concerning the role of proof, where significant effects are found, they are positive with respect to truth and negative for explanation. The truth role is only a significant predictor for the algebra question and has the effect of decreasing student propensity towards an incorrect empirical verification of the false conjecture. In the geometry question, students who felt proof has an explanatory function were more, not less, likely to select naive arguments, although this only reaches significance for two forms (narrative and counter-example). Overall, these observations add further support to finding F15*.

Student sex is associated with different choices in both questions: in algebra, girls were more likely than boys to prefer empirical (incorrect) to narrative (correct) arguments; in geometry, negative estimates are obtained for girls for the formal correct and narrative choices and positive estimates for formal incorrect and counter-example. Since the comparison options (empirical and naive) and the formal incorrect option all present an incorrect conclusion, these estimates suggest that girls were slightly more likely than boys to choose an approach which accepts rather than refutes a false conjecture. But in geometry, girls were more likely than boys to choose correctly a simple counter-example to refute the conjecture. Overall, we report:

F23. There are differences in the choices of argument in response to a false conjecture between girls and boys, and these differences are particularly marked in geometry.

Turning to the Level 2 variables, we found that in schools that are selective or are outside inner city areas, students chose the comparison category less frequently than others, and choices also differed between those from mixed or single-sex schools. Curriculum factors, and the department's overall approach to teaching proof, while associated with choices in both questions, do not have a consistent effect, and tend not to hold for all the arguments. Two variables do have consistent effects for all arguments: in geometry, there are consistent positive estimates for a school's selection procedure; in algebra, consistent negative estimates linked to following the London examination syllabus. We therefore report:

F24. Level 2 variables have significant influences on student responses to false conjectures, but only two have consistent effects: in geometry, students in

selected schools are more likely than others to choose a correct refutation, but also more likely to choose an incorrect formally-presented argument; and in algebra, students following the London examination syllabus are more likely than others to choose the incorrect empirical verification.

SCHOOL DIFFERENCES

After all the variables listed in Table 11 above are added to the models for these two questions, no significant variation is left at the school level, suggesting that student choices for these two questions did not differ significantly according to the school attended.

F25. There is no between-school variation in the arguments chosen in response to a false conjecture, after account is taken of all the variables that influence student choices.

6. SUMMARY: STUDENT RESPONSES TO THE MULTIPLE-CHOICE QUESTIONS

From our analyses of students' response patterns to the multiple-choice questions, we conclude that a multitude of factors need to be taken into consideration in order to interpret the choices that students make. The arguments students select are influenced by the criteria underlying choice: whether they are choosing as their own approach or for the best mark. When choosing an argument they think will receive the best mark, students are more likely to choose one that is mathematically valid than when they choose one that is closest to what they would do. This is largely because they believe that formal presentations will receive the best mark for direct proofs and refutations alike, in both algebra and geometry. Generally speaking, empirical verifications or refutations (i.e. the use of a counter-example) are not considered for the best mark in either algebra or geometry and students are more likely to select these forms as closest to their own approach. This result is particularly interesting for the refutations since an empirical form in this case is perfectly valid.

When students are asked to select which argument would be nearest to their own approach, in algebra the most popular approaches are those presented in prose-form and the least popular are formal. Formal forms are selected more frequently in geometry than in algebra. Our results show clearly that while students' mathematical background (as measured by their Key Stage 3 score) is an important factor in understanding their choices, many other factors have a role in students' preferences. Three student factors that are strongly and consistently related to students' choices are their assessments of the generality and explanatory power of a proof — they are likely to favour proofs when they know they are general, when they find them helpful in explaining the statement, and when they believe they will receive a good mark. When it comes to visual proofs, however, of these three factors only its explanatory power appears to play a role. Student choices are also related to their view of the role of proof, with those who have no sense of what proof is for most likely to choose empirical forms. Choices also vary according to student sex, although a general pattern to capture the different preferences is hard to identify.

Influences on choices are not limited to student factors, although curriculum factors are more prevalent in relation to algebra than geometry. This is perhaps not surprising as in the majority of schools, proof is addressed mainly through investigations, most of which have an algebraic content. We found that students who have more mathematics teaching than average are less likely to choose empirical arguments (with the converse also true). Most other curriculum factors however have rather specific influences which differ according to the mathematical content of the question. One interpretation of this is that, in general, curriculum variables have a conceptual influence affecting how much students know about a particular proof, but not a metacognitive one; that is they may not influence a student's conception of the process of proving.

Finally, after taking into account the various student, school, curriculum, teaching and teacher influences on choices, we find little variation in algebra choices according to the school attended but rather more difference in choices in geometry. We could argue that the lack of emphasis on geometry in the current National Curriculum leaves room for more variation in teaching than in algebra, with a few schools choosing to place a greater emphasis on the geometry domain than demanded.

7. STUDENT VIEWS OF THE GENERALITY OF A VALID PROOF

We were interested to see whether students understood the *generality of a valid proof* — that is, if they were aware that if a statement had already been proved over a particular domain, it also was true in any subset within its domain of validity. Two questions, one in algebra, A2, and one in geometry, G2, were included in the questionnaire to explore this; that is whether or not students saw proof as general. Figure 14 presents students' responses to these two questions.

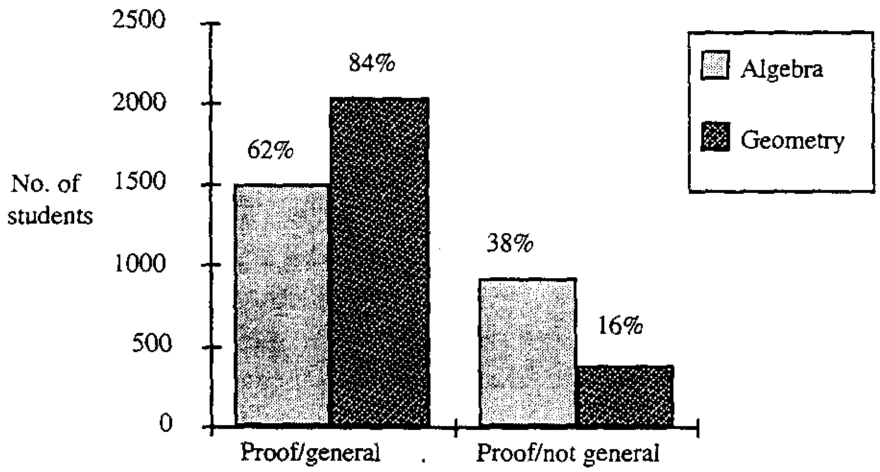


Figure 14: Distribution of students' responses to A2 and G2.

For both questions, we found that the majority of students were aware that once a statement had been proved in general no further work was necessary to check if it applied to a particular subset within the domain of validity. In contrast to all other questions, more students answered this question correctly in geometry (84%) than in algebra (62%), a surprising result to be followed up through interviews with selected students. In general, we report:

F26. The majority of students know that once a statement has been proved it holds for all cases within its domain of validity.

8. DESCRIPTIVE STATISTICS: STUDENTS' CONSTRUCTED PROOFS

Four questions required students to construct a proof of their own (A4, A7, G4 and G7). All students' constructed proofs were classified according to two criteria: a score for correctness (on a scale of 0 to 3 according to the criteria presented in Table 3, on page 5); and the main form of argument used, classified as none, naive, empirical, exhaustive, enactive, formal, narrative, visual or counter-example[24].

CONSTRUCTED PROOF SCORES

We started with the analysis of student scores for constructed proofs. The first step of analysis was to compare the distribution of student scores with the distribution of choices of correct proofs. For both algebra and geometry comparison of the overall proportions for all questions showed that a significantly higher proportion of students selected a correct proof than constructed either a partial or complete proof (algebra: $\chi^2 = 1088.77$, df = 1, $p < 0.0001$; geometry: $\chi^2 = 961.29$, df = 1, $p < 0.0001$), indicating:

F27. Students are better at choosing correct mathematical proofs than constructing them in both algebra and geometry.

We then looked at the distribution of students' constructed proof scores for each question. This information is presented below in Figures 15-18:

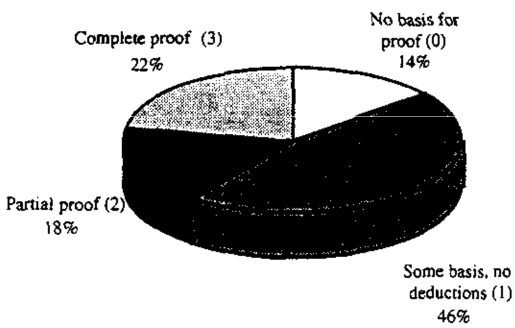


Figure 15: Distribution of scores in A4.
(mean = 1.471)

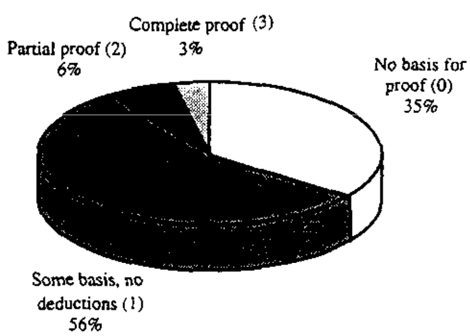


Figure 16: Distribution of scores in A7.
(mean = 0.778)

24 This last category was used for A7 only, where a number of students argued that 0 was not a multiple of 4 and therefore showed that the statement was false.

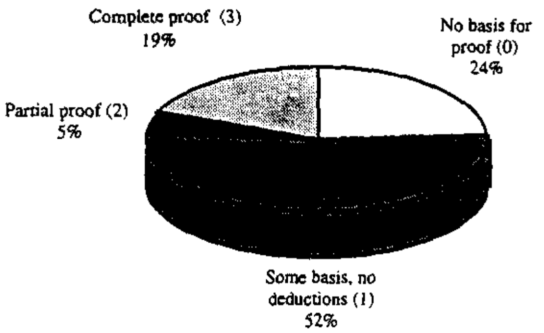


Figure 17: Distribution of scores in G4.
(mean = 1.188)

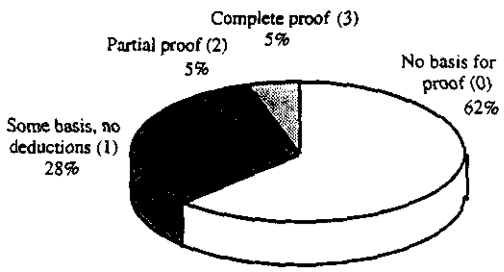


Figure 18: Distribution of scores in G7.
(mean = 0.522)

The distribution of scores shows a consistent pattern of poor student performance, even among our sample of high-attaining students. In no question was the average score greater than 1.5 (half of the maximum) and for the unfamiliar questions, A7 and G7, the means were well below 1. In G7, a large percentage (62%) could do nothing at all towards writing the proof.

It is all too clear that many students were not even able to begin to construct a proof and, if they did make a start, could only indicate some relevant information unconnected by logical reasoning. Deductive reasoning was shown by a percentage of students which varied according to the mathematical content of the question, with rather more displayed in the familiar algebra proof and very little in the unfamiliar questions (40% for A4; 9% for A7; 24% for G4; and 10% for G7). We therefore report:

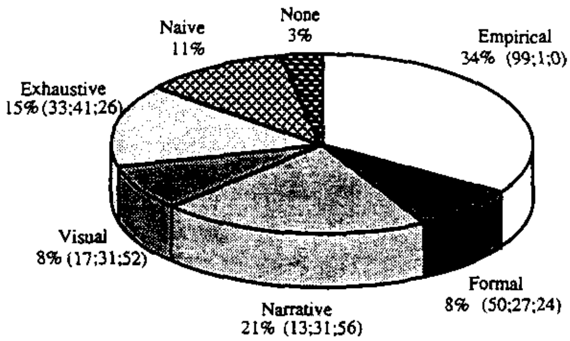
F28. Students are unlikely to use deductive reasoning when constructing their own proofs.

A comparison of both the mean scores in the two domains and the percentages of students using no deductive reasoning suggests that students had more success in constructing proofs in algebra than in comparable questions (in terms of familiarity) in geometry. We therefore compared student responses in A4 with those in G4, and A7 with those in G7. In both cases the differences in scores are significant (A4 and G4: $t = 9.96$, df = 4916, $p < 0.0001$; A7 and G7: $t = 11.91$, df = 4916, $p < 0.0001$). We conclude:

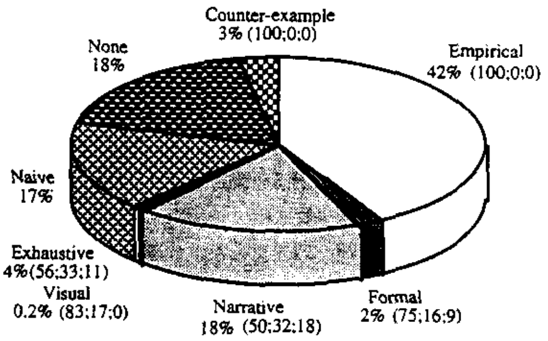
F29. Students are better at constructing proofs in algebra than in geometry.

FORM OF ARGUMENT IN CONSTRUCTED PROOFS

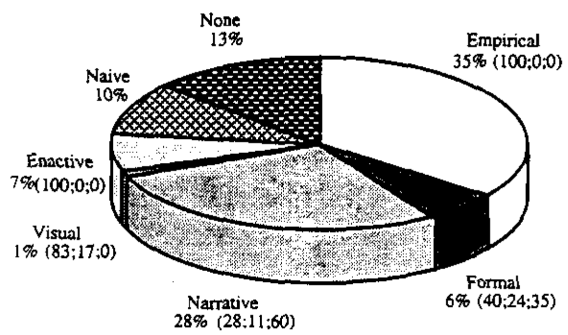
We now turn to the forms of argument students used in their attempts to construct proofs of their own. Student proofs often contained different forms of argument and initially each one was coded. A second level of coding was then undertaken to specify the main form used. If the argument consisted entirely of incorrect statements or simply restated the problem, then it was classified as naive and received a score of 0. Proofs classified as empirical were assigned a score of 1, unless there was evidence that the students were treating the example(s) they presented as generic, in which case they were given a score of 2. If students gave a counter-example to show that the statement in A7 was incorrect, they received a score of 1 because although they were wrong there was evidence of understanding the role of counter-example in the proving process. All other forms of argument received a score of 1, 2 or 3 according to how far they were logical and analytic. Figure 19 shows the overall percentages of each form of argument used in each question along with a breakdown for each argument of the distribution of scores.



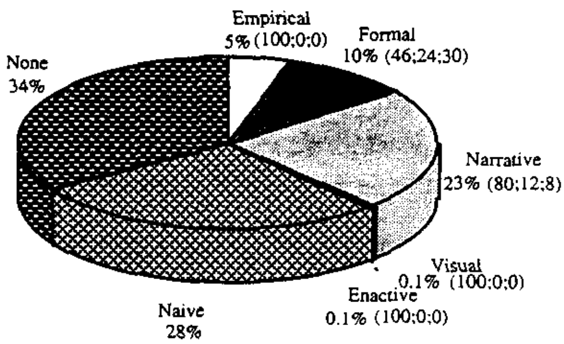
Arguments used in A4.



Arguments used in A7.



Arguments used in G4.



Arguments used in G7.

Figure 19: Distribution of forms of argument used in constructed proof.
(% of arguments obtaining a score of 1, 2 or 3 in brackets)

Figure 19 shows that in all 4 constructed proof questions, more students presented their arguments in narrative than formal forms (A4: 21% narrative, 8% formal; A7: 18% narrative, 2.2% formal; G4: 28% narrative, 6% formal; A4: 28% narrative, 10% formal).

F30. Students construct narrative arguments more frequently than formal arguments.

Additionally, in general, a higher proportion of narrative arguments than formal ones (or indeed any other form) were correct, except for unfamiliar geometry questions where the reverse was the case.

F31. Narrative arguments are more likely to be completely correct than other arguments, except when students attempt to prove an unfamiliar geometry statement.

We also notice that an empirical argument was the most popular in all but the unfamiliar geometry question (A4, 34%; A7, 42%; G4, 35% and G7, only 5%), and that for this question, the majority of students constructed either no argument at all (34%) or one classified as naive (28%).

F32. Empirical verification is the most popular form of argument used in constructing proofs, except in the case of an unfamiliar geometry proof.

In view of our finding noted previously that an empirical argument was not necessarily the most popular for a student's own approach, we tested to see if the distribution of empirical choices was the same as the distribution of empirical constructions. We found significant differences in algebra ($\chi^2 = 246.73$, df = 1, $p < 0.0001$), but not geometry ($\chi^2 = 1.94$, df = 1, NS).

F33. In algebra but not geometry, students are more likely to construct empirical arguments than to choose them.

Finally, we noted the presence of a larger than expected proportion of a particular form of argument in a constructed proof when a similar type of proof had already been presented in the preceding multiple-choice question. For example, in algebra, exhaustive and visual proof forms were presented in A1 but not in A6. Figure 19 shows that 15% of students used exhaustive arguments to prove the statement in A4, compared to only 0.4% in answer to A7; for the visual form, the percentages are 8% and 0.2% respectively. For the geometry questions, a similar pattern is evident for the enactive form, used by 7% of students for G4 and only 0.1% for G7. The pattern holds for visual arguments too, although these were rare for both questions (1% and 0.1% respectively). These data suggest that some students may be able to adapt proofs previously shown to them in order to construct their own proofs. The contextual factors that surround this productive strategy will be investigated in follow-up student interviews.

9. MULTILEVEL MODELLING: STUDENTS' CONSTRUCTED PROOFS

The next stage in the analysis of responses to the constructed proof questions was to identify which, if any, of the Level 1 and Level 2 variables are associated with student constructed proof scores and to determine whether these influences vary from school to school. Multilevel models of student scores on each of the four questions, A4, G4, A7 and G7 were constructed[25] and multinomial models constructed to examine the forms of argument adopted. In the next section, we present the first set of models and the findings related to the proof scores.

MODELS OF CONSTRUCTED PROOF SCORES

Table 13 presents the estimated effects of the significant variables on the scores in the constructed proof questions. To explain how these estimates can be interpreted, we focus on the model for the familiar algebra proof, A4. The model shows that three Level 1 variables and one Level 2 variable are associated with students' scores: student sex; their

25 The procedures for modelling continuous output variables are described in Appendix 4.

Key Stage 3 test score; their view of algebra proofs as general or not; and the percentage of the class expected to be entered in the higher-tier GCSE paper. Hence, the base group for this model was selected by according these variables particular values: male students with a KS3 score of 6, who were from a class where 80% of students were expected to be entered for the GCSE higher-tier paper and who believed a valid algebra proof not to be general. The estimate for this *base group mean* was 1.28.

The estimates for the explanatory variables indicate the expected increase (or decrease) in this mean score. For example, to calculate the estimated score for a female student with a KS3 score of 8, who was aware of the generality of a valid algebra proof and who came from a class where all students were expected to be entered for the higher paper, we would add to the base group mean of 1.28, an additional 0.13 for the effect of being female, 2 * 0.29 for the Key Stage 3 effect, 0.20 as the estimate associated with the variable, proof as general, and 20 * 0.003 to take into account that 100% of the class entered GCSE higher tier. The estimated score for this group would therefore be 2.25.

Also in Table 13 (in italics) are the standardised effects associated with each explanatory variable, so that their relative effects within a model can be compared. On question A4, for instance, the standardised estimates indicate that the variable with the largest effect is Key Stage 3 test score (0.207), the smallest is a student's sex (0.067), while the other two significant variables have similar effect sizes; 0.098 for recognition of proof as general and 0.089 for % GCSE higher tier.

	Algebra constructed proofs		Geometry constructed proofs	
	A4	A7	G4	G7
Base group mean	1.28 (0.053)	0.62 (0.041)	0.803 (0.057)	0.39 (0.034)
Variables				
Level 1 Views of role of proof Truth		0.075 (0.029) *0.052*	0.140 (0.040) *0.068*	
Student characteristics Sex	0.13 (0.053) *0.067*	0.07 (0.028) *0.049*		
KS3 test score	0.29 (0.034) *0.207*	0.15 (0.023) 0.150	0.319 (0.033) *0.225*	0.29 (0.034) *0.257*
Responses to questionnaire Proof as general (algebra)	0.20 (0.040) *0.098*			
Proof as general (geometry)			0.256 (0.044) *0.175*	0.24 (0.032) *0.109*
Level 2 Curriculum factors % GCSE higher tier	0.003 (0.001) *0.089*			0.0018 (.0006) *0.066*
Main textbook SMP		0.10 (0.050) *0.067*		
Vickers		0.21 (0.071) 0.100		
Rayner		0.16 (0.076) *0.063*		
Approaches to teaching proof Write geometry proofs			0.181 (0.060) *0.089*	
Notes: Standard errors shown in brackets; standardised effects shown in italics.				

Table 13: Estimated effects of the significant variables on the scores for the four constructed proof questions.

Clearly Key Stage 3 score has a considerable influence, not only by the size of its effect in comparison with other variables, but also because its effect spans all four questions. We therefore report:

F34. Students with higher Key Stage 3 test scores are better at constructing proofs than those with lower scores.

Student sex is a significant factor in algebra, with girls predicted to obtain higher scores than boys on both questions, leading us to report:

F35. In algebra, girls construct better proofs than boys.

The other student factors associated with how well students were able to construct proofs concern their view of a proof, and their appreciation of its generality. An awareness of the generality of a valid algebra proof is associated with an improved score in A4, (0.098). Similarly, recognising the generality of a valid geometry proof meant students were likely to construct better proofs for both the geometry questions (G4, 0.175; G7, 0.109). Furthermore, students who recognised that proof had a role in verifying the truth of a statement outperformed those who did not, at least on the unfamiliar algebra question and the familiar geometry one (A7, 0.052; G4, 0.068). We report:

F36. Students who recognise the generality of a valid proof and appreciate its role in establishing the truth of a statement are better at constructing proofs than those who do not.

Only one Level 2 variable in each model has a significant effect: in A4 and G7, the variable is % GCSE higher tier; in A7, it is main textbook; and, in G4, it is a teaching approach in which students are expected to write geometry proofs. We note that these factors all relate to curriculum and teaching issues, rather than school or teacher variables.

F37. At least one curriculum or teaching factor is associated with students' constructed proof scores: variables of significance include the percentage of students in the class entering the GCSE higher-tier examination, the textbook followed and the expectation that students will write geometry proofs.

SCHOOL DIFFERENCES

We now turn to look at the variation in response between students from different schools; that is, to test if there were schools in which students obtained rather higher (or lower) scores than would be expected after adjustment according to the significant variables in our models. Table 14 below presents the random effects associated with the variance component[26] model for each constructed proof score. Level 1 and Level 2 variations represent respectively estimates of the student and school deviations from the means as predicted by the fixed part of the model. Intra-school correlation measures the proportion of the total variation which is between-schools.

	Constructed proof questions			
Random effects	A4	A7	G4	G7
Level 1 variation	0.86 (0.026)	0.46 (0.026)	0.85 (0.026)	0.52 (0.016)
Level 2 variation	0.04 (0.011)	0.03 (0.006)	0.05 (0.013)	0.05 (0.011)
Intra-school correlations	4.0%	5.8%	5.6%	8.7%
Notes: Standard errors in brackets.				

Table 14: Random effects for models of constructed proof scores.

Table 14 shows that for all constructed proof scores, there was substantially more variation in the performances of students *within* schools than *between* schools. In Figure 20, 95% uncertainty intervals around the residual estimates for each school have been plotted.

26 Underlying a variance component model is the assumption that schools vary around their predicted means in the same way — i.e. that the effect of a given explanatory variable will be the same in all cases.

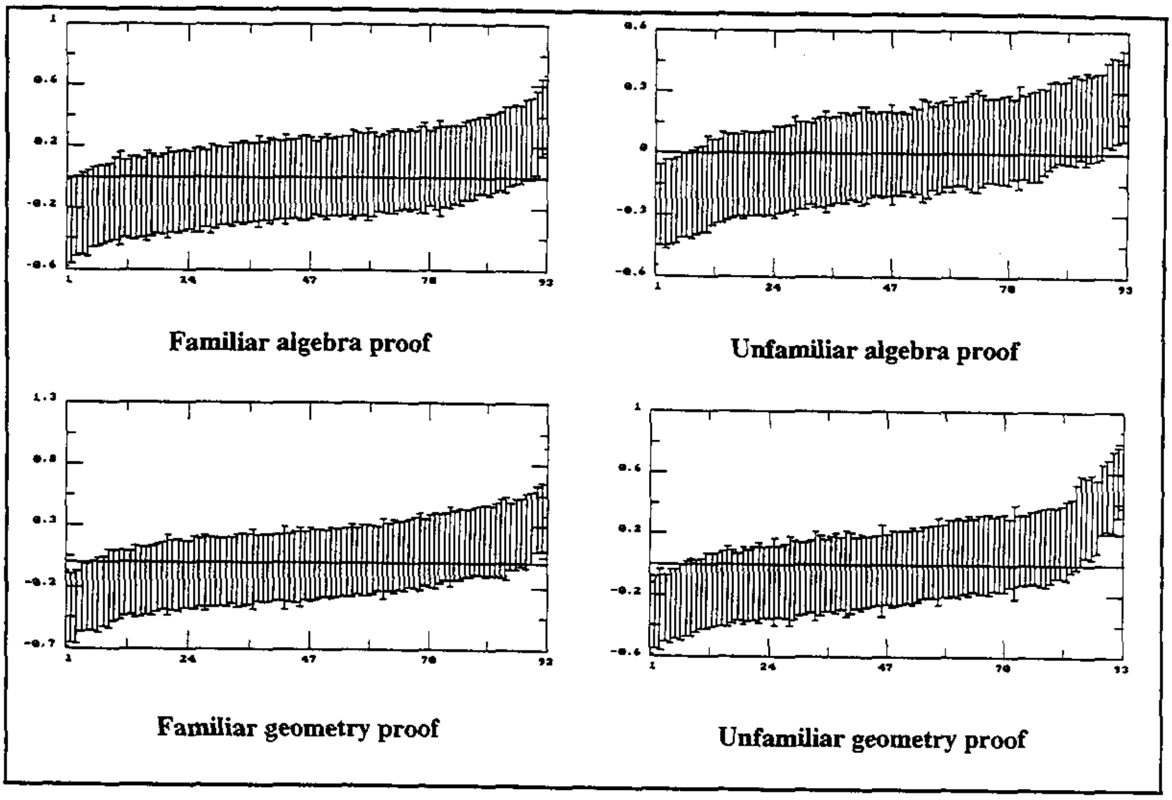


Figure 20: School deviations from predicted means: 95% uncertainty intervals around estimates for each school.

It is clear from Figure 20 that, after adjusting for the significant variables, there was considerable overlap between schools. We are most interested in the schools at the upper and lower extremes of the plots, since it is in these that students were obtaining scores which differed most from what would be expected from our models. As explained previously, to investigate the reasons for these between-school differences, we selected from each plot the five top and bottom schools to add to the sample from which case study schools would be selected in the second phase of our research.

In Table 14, the highest intra-school correlation of schools is obtained for the scores on the unfamiliar geometry question, G7, indicating that a larger proportion of the total unexplained variance was between-schools for this question than for the others. If we also look at the plot of school residuals associated with student scores on G7 (Figure 20), there is a rather sharper increase in the gradient at the upper extremes, suggesting that, in a handful of schools, students were performing especially well. It is interesting to note that this pattern mirrors that identified by the multinomial modelling, where unexplained school variation was similarly largest amongst the responses to the unfamiliar geometry question.

F38. Although there is little between-school variation in students' scores for constructed proofs, there are some schools whose students score better or worse than predicted.

We also fitted more complex random coefficient models[27] to each of the four scores to explore whether or not the effects of explanatory variables were the same across all scores. Significant effects are found for two scores only: the scores for the familiar algebra proof (A4) where the effect associated with student sex varies significantly at the school level (see Table 15); and the unfamiliar geometry question where the effects of Key Stage 3 score are not the same in all schools (see Table 16).

	Constructed proof A4
Level 1 Student-level variance	0.84 (0.026)
Level 2 Variance of boys around school means for boys.	0.10 (0.026)
Covariance of boys/girls	0.014 (0.014) *0.273*
Variance of girls around school means for girls.	0.026 (0.13)
Notes: a. Standard errors in brackets. b. Correlation coefficient in italics.	

Table 15: Random effects in constructed proof scores on A4.

The estimated random effects for A4 presented in Table 16, indicate that boys' performances varied according to the school attended considerably more than those of girls. The estimate for girls is very small (0.026) and barely reaches significance, but the estimate of 0.10 for boys suggests that, in some schools, boys were obtaining scores up to 0.62[28] above or below the predicted mean. The positive estimate for covariance, (0.014), suggests that in schools where the boys' scores were better than predicted, the same was also true for the girls. However, the correlation coefficient is very small, indicating that this pattern was not significant. We therefore report:

F39. Schools make a significant difference to how well boys construct familiar algebra proofs, although this is not the case for girls.

	Constructed proof G7
Level 1 Level 1 variance	0.49 (0.014)
Level 2 Variance around the school estimate	0.45 (0.0098)
Covariance of school estimate/KS3 estimate	0.035 (0.0094) *0.813*
Variance around KS3	0.42 (0.012)
Notes: a. Standard errors in brackets. b. Correlation coefficient in italics.	

Table 16: Random effects in constructed proof scores in G7.

In G7, we also found that schools affected students' proofs in different ways, although in this case, they varied according to Key Stage 3 score. The estimates presented in Table

27 Details of random coefficient models can be found in Appendix 5.
28 1.96 multiplied by the square root of the variance.

16 indicate that, not only did schools vary around their predicted means (0.45), but also that the effects associated with Key Stage 3 score differed from school to school (0.42). A positive estimate for the covariance (0.035) is obtained, and the large correlation coefficient (0.813) indicates that in schools where student performances were better than the fixed model would predict, the size of the Key Stage 3 effect differed between schools i.e. where students perform well in general, students with higher Key Stage 3 scores do particularly well.

F40. In schools with better than predicted performance in constructing unfamiliar geometry proofs, students with high Key Stage 3 scores do particularly well.

MODELS OF FORMS OF ARGUMENT IN CONSTRUCTED PROOFS

In addition to constructing models of scores for constructed proofs, we were also interested in the factors related to the forms of argument used in the proofs. As with the multiple-choice questions, multinomial models of the categorical output were constructed, one for each of the four constructed proofs. From these models, variables significantly associated with responses could be isolated and their effects in relation to students' preferences for particular forms of proof identified. The comparison category chosen for these multinomial models was the group of students who produced little if anything in terms of a proof. To simplify the models and obtain more reliable estimates, some other categories were grouped together[29].

Table 17 below lists the χ^2 values of all variables significantly associated with the form of argument used by students in each of their four constructed proofs.

29 In A7, visual and exhaustive arguments were included in narrative and, in G7, visual and enactive arguments were included in narrative.

Variables	Constructed proof questions A4	A7	G4	G7
Level 1				
Views of role of proof				
Truth		43.61***(4)	180.2***(5)	
Explanation		13.82** (4)		
Student characteristics				
Sex	653.4*** (5)	92.83*** (4)		
KS3 score	262.3*** (5)	119.6*** (3)	515.3*** (5)	290.9*** (3)
Level 2				
School factors				
Location	12.60* (5)			12.54** (3)
School sex				102.4*** (6)
Selection procedures			12.95* (5)	32.77*** (3)
Curriculum factors				
Examination syllabus	25.78* (15)			31.22** (9)
Main textbook/ scheme		35.25** (16)	33.21* (20)	36.71*** (12)
Hours of mathematics		207.4*** (4)		
Approaches to teaching proof				
Write geometry proofs				20.24*** (3)
Teachers' views of National Curriculum				
Emphasis on formal proof		44.75*** (8)		

Notes:
a. * = p<0.01; ** = p<0.001; ***= p<0.0001.
b. df shown in brackets.

Table 17: χ^2 values showing the effects of significant variables associated with the form of argument used in constructed proofs.

Table 17 shows that only Key Stage 3 score is associated with the form of argument used in all four constructed proofs.

F41. Key Stage 3 test score influences the form of argument used in a proof.

Students' views of the role of proof also had an effect on the argument they used, although not in all questions. Additionally in algebra proofs, girls and boys adopted significantly different forms of argument.

F42. Boys and girls use different forms of argument in algebra.

Table 17 shows that at least one Level 2 variable is significant, the most consistently influential being the main textbook used.

F43. The form of argument used in constructing a proof is associated with at least one Level 2 variable, with the main textbook used having the most consistent effect.

Tables 18 and 19 below present the estimated effects of the significant variables on the form of proof used in algebra and in geometry proofs respectively.

		Forms of argument					
		Empirical	Exhaustive	Formal	Narrative	Visual	Counter-example
Base group ratio	A4	1.17 (0.26)	-0.36 (0.30)	-1.34 (0.41)	-0.11 (0.25)	-1.69 (0.41)	—
	A7	-0.26 (0.48)	—	-2.92 (0.62)	-1.90 (0.25)	—	n/a
Variables							
Level 1							
Views of role of proof							
Truth	A7	0.24 (0.09)	—		0.52 (0.12)	—	
Explanation	A7		—	-1.31 (0.50)	0.31 (0.11)	—	
Student characteristics							
Sex	A4	0.72 (0.09)	0.42 (0.12)	0.37 (0.16)	0.69 (0.10)	0.56 (0.15)	—
	A7	0.50 (0.09)	—		0.52 (0.12)	—	
KS3 score	A4	-0.36 (0.07)	0.24 (0.09)	0.83 (0.13)	0.71 (0.08)	0.60 (0.13)	—
	A7		—	1.21 (0.26)	0.80 (0.09)	—	0.68 (0.18)
Level 2							
School factors							
Location	A4				0.42 (0.18)	0.84 (0.28)	—
Curriculum factors							
Examination syllabus							
SEG	A4				-0.78 (0.26)		—
London	A4	-0.81 (0.30)			-0.75 (0.28)		—
Main textbook							
SMP	A7	0.43 (0.18)	—			—	
Vickers	A7	0.74 (0.26)	—			—	
Rayner	A7	0.92 (0.27)	—			—	
Hours of mathematics	A7					—	13.2 (0.86)
Teachers' views of National Curriculum							
Emphasis on formal proof	A7		—		-1.53 (0.76)		-4.62 (0.26)

Notes:
a. The comparison category consists of students who constructed no or naive arguments.
b. Use of visual and exhaustive arguments modelled as separate categories in A4 only; and counter-example applies only to A7.
c. Standard errors in brackets.
d. — indicates category not modelled for this question.
e. n/a indicates the number of students in the base group choosing this option was too small to obtain an estimate.

Table 18: The estimated effects of the significant variables on the forms of argument used in algebra proofs.

		Forms of argument				
		Empirical	Enactive	Formal	Narrative	Visual
Base group ratios	G4	-0.23 (0.23)	-2.05 (0.34)	-2.58 (0.33)	-0.74 (0.24)	-4.37 (0.72)
	G7	-3.11 (0.65)		n/a	-1.08 (0.83)	
Variables						
Level 1						
Truth	G4	0.31 (0.10)	0.45 (0.17)	0.79 (0.19)	0.47 (0.10)	1.13 (0.58)
KS3 score	G4		0.58 (0.13)	1.14 (0.14)	0.95 (0.08)	0.40 (0.37)
	G7	-0.43 (0.16)		1.44 (0.12)	0.77 (0.09)	
Level 2						
School factors						
Location	G7	0.64 (0.27)			0.35 (0.15)	
School sex						
Girl-only	G7				-1.01 (0.39)	
Boy-only	G7			-1.52 (0.19)	-0.87 (0.23)	
Selection procedures	G4	0.53 (0.24)	-0.74 (0.37)	-0.77 (0.34)		
	G7			-0.71 (0.14)		
Curriculum factors						
Examination syllabus						
SEG	G7			-0.65 (0.22)		
MEG	G7			-1.15 (0.27)		
London	G7			-0.54 (0.24)		
Main textbook/scheme						
SMP	G7			-0.48 (0.21)		
Holderness	G4			-0.88 (0.41)		
	G7			-0.33 (0.11)		
Vickers	G7			-0.94 (0.25)		
Rayner	G7			-1.20 (0.36)		
Approaches to teaching proof						
Write geometry proofs	G7			0.59 (0.16)		

Notes:
a. The comparison category consists of students who constructed no or naive arguments.
b. Use of visual and enactive arguments modelled as a separate category in G4 only.
c. Standard errors in brackets.
d. — indicates category not modelled for this question.
e. n/a indicates the number of students in the base group choosing this option was too small to obtain an estimate.

Table 19: The estimated effects of the significant variables on the forms of argument used in geometry proofs.

Tables 18 and 19 show that significant and positive estimates are consistently produced for Key Stage 3 score in relation to all forms except the empirical: for the unfamiliar algebra and the familiar geometry proofs, Key Stage 3 score has no significant effect on the extent to which students constructed empirical arguments; for the other two proofs, negative estimates are obtained (A4, -0.36 and G7, -0.43), suggesting that students with higher Key Stage 3 scores were more likely not to construct anything or to produce a naive argument rather than use an empirical argument. We therefore rewrite finding F41 below:

F41. With increase in Key Stage 3 test score, students become more likely to construct an argument with some relevant information, provided this is not empirical.

Significant effects are found for the variable, truth, in both the unfamiliar algebra and the familiar geometry proofs and were always positive. This suggests that, in these situations, students who believed proof has a role in establishing truth constructed something better than a naive argument more often than students who did not mention this role, although their preferred form of argument varied according to the question.

Significant effects for the variable, explanation, are found in the unfamiliar algebra question only and these suggest that regarding proof as having an explanatory function decreased the likelihood of constructing a formal argument and increased the likelihood of constructing a narrative one.

F44. When attempting to construct an unfamiliar algebra proof, students who believe that proof has an explanatory function are less likely than those who do not, to construct formal arguments and more likely to construct narrative ones.

The estimates presented in Table 18 for the student sex effect are all positive, indicating that girls were more likely than boys to come up with some kind of argument (other than the naive) for both of the algebra proofs. This finding reflects the difference in scores on the algebra proofs and tells us that more girls than boys achieved a score greater than zero.

Turning to the Level 2 variables, we find a general inconsistency in their effects across questions[30], with only one observation worth reporting:

F45. Attending a school in which students are expected to write geometry proofs increases the likelihood that students will construct a formal argument when tackling an unfamiliar geometry proof.

SCHOOL DIFFERENCES

Next we looked at the variation in form of argument between students from different schools. In Table 20 below, estimates of the amount by which schools varied around their predicted ratios for each form or argument in algebra, and the covariances between the different forms of argument are presented.

30 As was the case for the multiple-choice questions, most findings are specific to particular questions and argument forms. For example, for the familiar algebra proof and with respect to the main textbook we find: students following SMP, Vickers or Rayner are more likely than students using other books or schemes to construct empirical arguments rather than a naive proof or no proof at all.

		Empirical	Exhaustive	Formal	Narrative	Visual
Empirical	A4	**0.53 (0.11)**				
	A7	**0.28 (0.07)**	—			
Exhaustive	A4	0.43 (0.09) *.82*	**0.51 (0.12)**			
	A7	—	—			
Formal	A4	0.27 (0.11) *.44*	0.28 (0.12) *.47*	**0.71 (0.20)**		
	A7	0.09 (0.14) *.18*	—	**0.91 (0.49)NS**		
Narrative	A4	0.34 (0.07) *.80*	0.39 (0.08) *.94*	0.23 (0.10) *.47*	**0.34 (0.09)**	
	A7	0.33 (0.07) *.83*	—	0.29 (0.19) *.40*	**0.56 (0.13)**	
Visual	A4	0.18 (0.10) *.23*	0.28 (0.11) *.50*	-0.01(0.09) *-.01*	0.27 (0.09) *.59*	**0.62 (0.18)**
	A7	—	—	—	—	—

Notes:
a. Exhaustive and Visual arguments modelled as a separate category in A4 only.
b. Correlation coefficients shown in italics.
c. NS indicates variation is not significant.
d. — indicates category not modelled for this question.

Table 20: Random effects (variance-covariance estimates) at school level for forms of argument used in algebra.

The estimates for the two algebra proofs are given in Table 20. Those presented in bold on the leading diagonal for the familiar algebra proof (A4) indicate that, even after taking into account all the significant variables, the extent to which students constructed particular forms of argument varied significantly according to school attended. Schools varied most in the use of formal arguments (0.71) and least around the usage of narrative proof forms (0.34). Figure 21 shows the 95% uncertainty intervals around the residual estimates for each school which indicate how far they deviated from the ratio predicted by the fixed part of the model. On these plots there is rather less overlap between schools than in previous plots, suggesting that schools may have more effect on the forms of argument their students adopt than on the other output measures modelled. However, the familiar increase in the gradient of the estimates is again evident and we selected from each plot the five schools with the largest positive estimates to add to our case study sample.

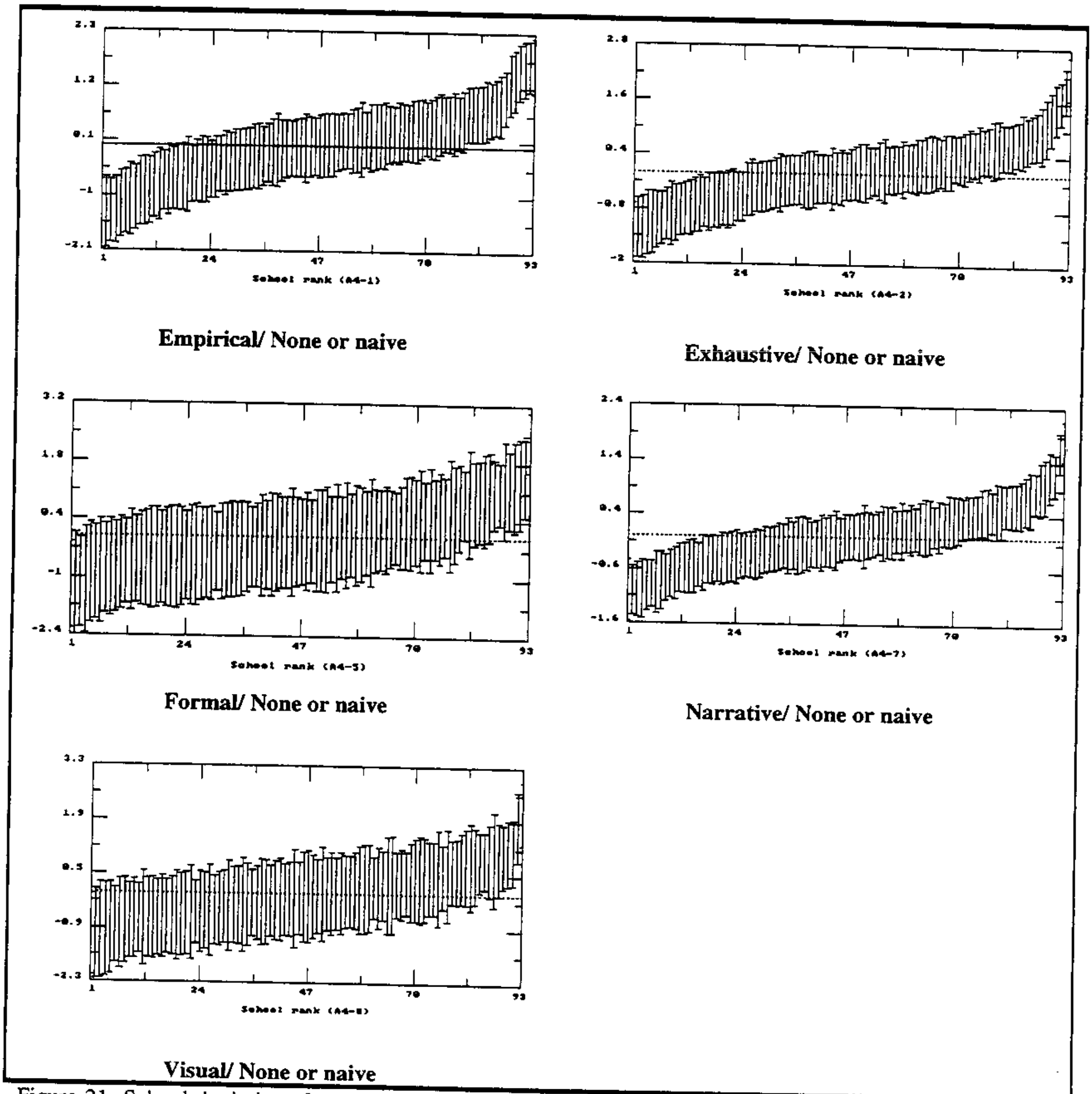
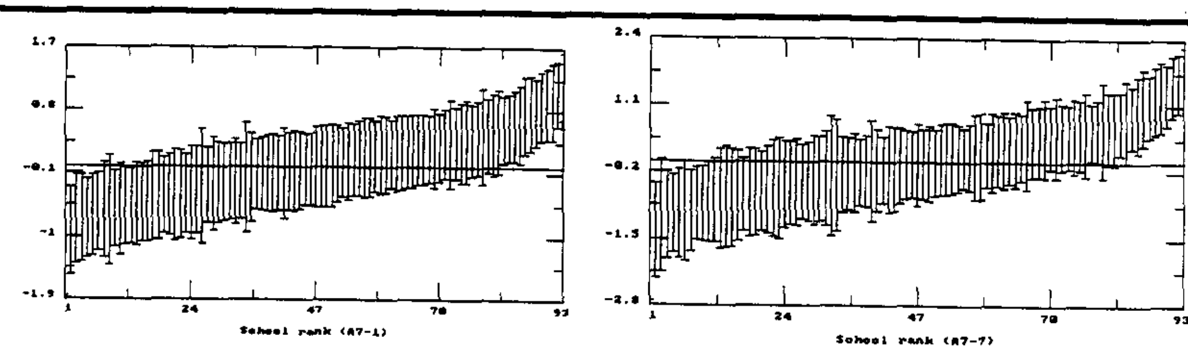


Figure 21: School deviations from predicted ratios: 95% uncertainty intervals around estimates for each school (A4).

The covariance estimates indicate that for A4 the strongest association is between the exhaustive and narrative forms (0.94), suggesting that, in schools where high proportions of students constructed exhaustive arguments, the proportion of narrative arguments constructed was high. Similar associations, although not as strong, are apparent between empirical argument and both exhaustive and narrative constructed proofs (0.82 and 0.80). Weaker correlations are found between visual arguments and the exhaustive and narrative forms (0.50 and 0.59). Associations between the formal and other forms are rather weak (exhaustive 0.47, narrative 0.47) or not correlated at all. On the whole, in schools where the ratio of students constructing formal argument to those producing no argument or a naive one was higher than usual, we would not expect higher proportions of all other forms as well.

For the unfamiliar algebra proof, the highest estimate for school variation is again obtained for the formal form. However, the standard error is very large (not surprising since so few students overall adopted this form) and the estimated difference not therefore significant. Significant estimates of school variation are obtained for both narrative (0.56) and empirical (0.28) arguments and a strong association evident between these two forms, indicating that in schools where narrative arguments were popular, so too were empirical verifications. Like the previous algebra constructed proof, weak associations or no correlations at all are found between the formal and other forms. In Figure 22 below, we present the 95% uncertainty intervals around the school residuals for the two forms of argument for which significant variation is found, and again we selected the top five from each to add to our sample of case study schools.



Empirical/ None or naive **Narrative/ None or naive**

Figure 22: School deviations from predicted ratios: 95% uncertainty intervals around estimates for each school (A7).

In Table 21, estimates of the amount by which schools varied around their predicted ratios for each form of argument in geometry, and the covariances between them are presented.

		Empirical	Enactive	Formal	Narrative	Visual
Empirical	G4	**0.59 (0.12)**				
	G7	**0.21 (0.15) NS**				
Enactive	G4	0.56 (0.12).81	**0.82 (0.21)**			
	G7	—	—			
Formal	G4	0.29 (0.11).53	0.42 (0.15).65	**0.51 (0.18)**		
	G7	n/a	n/a	**n/a**		
Narrative	G4	0.35 (0.09).57	0.63 (0.13).88	0.44 (0.11).78	**0.63 (0.13)**	
	G7	0.15 (0.07).98	—	n/a	**0.11 (0.06) NS**	
Visual	G4	0.07 (0.24).10	0.36 (0.34).44	0.41 (0.30).64	0.48 (0.25).67	**0.81(0.81)NS**
	G7	—	—	—	—	—

Notes:
a. Enactive and Visual arguments modelled as a separate category in G4 only.
b. Correlation coefficients shown in italics.
c. NS indicates variation is not significant.
d. — indicates category not modelled for this question.
e. n/a indicates too few responses for an estimate to be obtained.

Table 21: Random effects (variance-covariance estimates) at school level for forms of argument used in geometry.

For the familiar geometry proof (G4), significant school variation is found for all forms, except the visual (recall that only 1% of students overall actually constructed visual arguments) with schools varying most around their predicted ratio for enactive as compared to naive or no argument (0.82). Figure 23 illustrates the 95% uncertainty intervals around the residual estimates for each school.

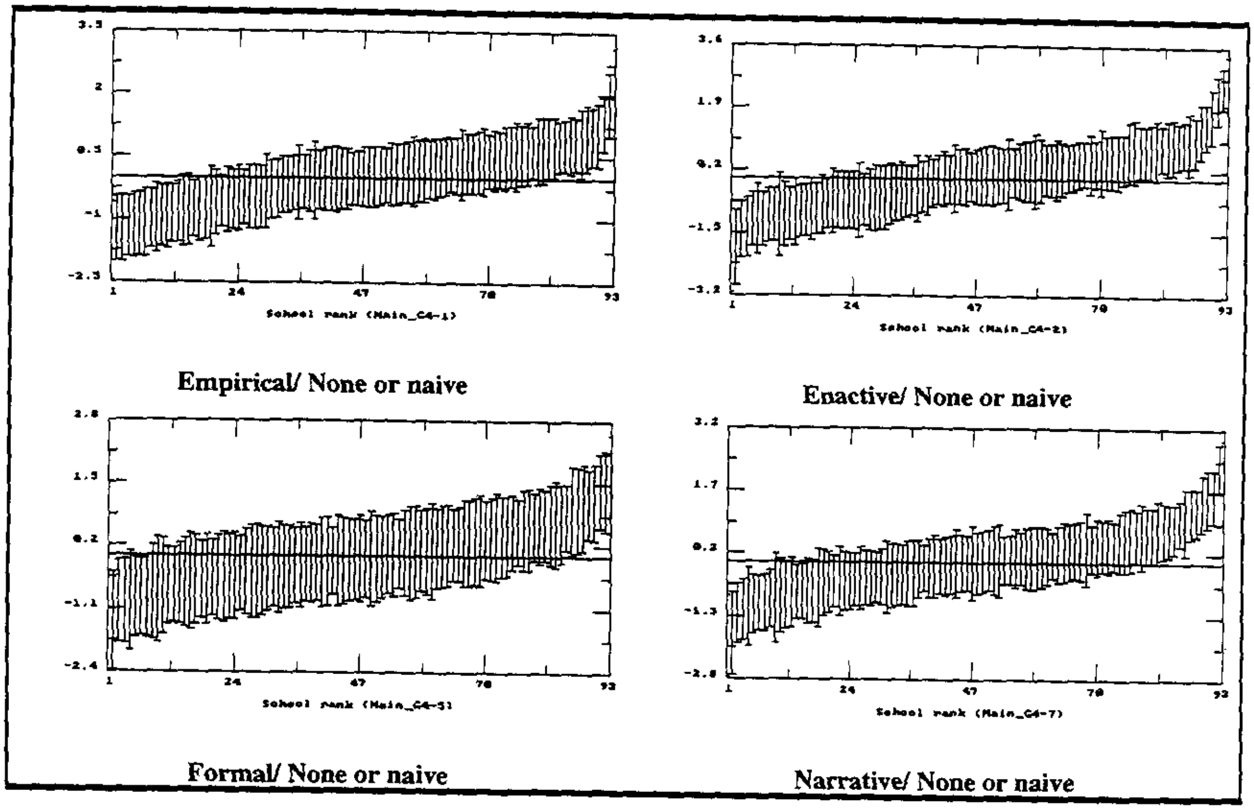


Figure 23: School deviations from predicted ratios: 95% uncertainty intervals around estimates for each school (G4).

The covariance estimates indicate that, for G4, there are positive associations between all forms of argument (although the association between empirical and visual is not significant). This indicates that in schools where only a small proportion of students were unable to do anything at all, higher numbers of all other forms of argument were produced, with the highest correlation between the use of enactive and narrative arguments (0.88). This finding may be related to the particular mathematical content of this question and reflect the multitude of different methods that can be used to demonstrate that the sum of the angles of a quadrilateral is 360°.

For the unfamiliar geometry proof, (G7), there is no significant variation at the school level, with no estimates obtained for the formal choice and very small estimates with large standard errors for the others. At first glance, this might seem surprising, since we know that scores varied considerably between schools for this proof. Our results suggest that the ways students tried to write their proofs were similar, regardless of school attended, although their success in constructing a valid proof was not.

F46. In general, the forms of argument used when students attempt to construct a mathematical proof differ according to school attended.

10. SUMMARY: STUDENTS' CONSTRUCTED PROOFS

It is clear from our analyses that our students do not find it easy to construct valid proofs. Despite being attracted by forms of argument they believe to be general and to have explanatory power, and despite thinking that formal arguments will gain them good marks, the majority of students do not incorporate deductive reasoning in their constructed proofs and very few even attempt to construct a formal argument. In fact, students are rather better at choosing correct mathematical proofs than constructing them in both algebra and geometry. As with the multiple-choice questions, students' responses differ according to whether they are working on algebra or geometry problems as well as on the specific mathematical content of the question — in particular, they obtain higher scores for their constructed proofs in algebra than in geometry and where the mathematical content of the proof is more familiar.

Whereas empirical verifications tend not to be the most popular when students are choosing proofs, they are the most frequently constructed arguments for all but the unfamiliar geometry question. In algebra especially, students are more likely to construct empirical arguments than to choose them. Where they are not able to construct an empirical argument, naive arguments become more frequent. Those students who construct neither empirical arguments nor naive arguments are likely instead to present their proofs in narrative forms. Formal arguments are adopted by only a very small minority. One interesting observation to note is that a sizeable minority of students appear to adapt proofs shown in other parts of the questionnaire in order to construct their own proofs. This is perhaps a promising sign — some students are at least able to follow and reconstruct logical arguments.

As with the multiple-choice questions, student factors influence the kind of proofs students construct, both in terms of their generality and the form of argument used. For proof scores, Key Stage 3 test score has the strongest influence of all the explanatory factors and the lower this score the more likely is the student to rely upon empirical evidence only. Girls are better at constructing algebra proofs than boys, although this does not appear to signify a higher degree of deductive reasoning amongst girls, only that boys more often give no argument at all or construct one that is naive. Students' views of the generality and role of proof also have an effect on how well they are able to construct proofs and, in the case of the unfamiliar algebra question, holding a view that proof has an explanatory role increases the likelihood that a student will employ a narrative argument, and decreases the likelihood of a formal argument. School, curriculum and teaching factors are related to students' performances on the constructed proof questions: for constructing a valid proof, it seems to be important to be in a class where most students will be entered for the higher-tier GCSE paper, while it is the main textbook or mathematics scheme used in the school which seems to have the most consistent influence on the form of argument used.

Not all of the variation in student responses can be explained by our student, school, curriculum and teaching factors and, for all the scores, there are differences in student performance according to the school attended. This is particularly true for the scores in the unfamiliar geometry question, where in a small number of schools, students do much better than predicted. For some scores, not all students respond in the same way to school influences: when it comes to constructing a familiar algebra proof, school differences are much greater for boys than for girls; and in schools where students are better than predicted at constructing unfamiliar geometry proofs those with higher Key Stage 3 score do especially well. Finally, schools also seem to influence the way in which students present the proofs they construct, in all cases except in answering the unfamiliar geometry question.

11. DESCRIPTIVE STATISTICS: STUDENTS' VALIDITY SCORES

We investigated the extent to which students appreciated the scope of validity of an argument in algebra and in geometry by examining the two validity scores, (AVS and GVS), both of which had a range of 0 to 20. (Details of how these scores were constructed can be found in section 2.) The mean score in algebra is 10.6, and in geometry is 6.65. This difference is highly significant ($t = 34.47$, df = 4916, $p < 0.0001$), showing that students are much better able to appreciate the scope of validity of an argument in algebra than in geometry.

The frequency distributions of the validity scores in algebra and geometry are shown in Figure 24.

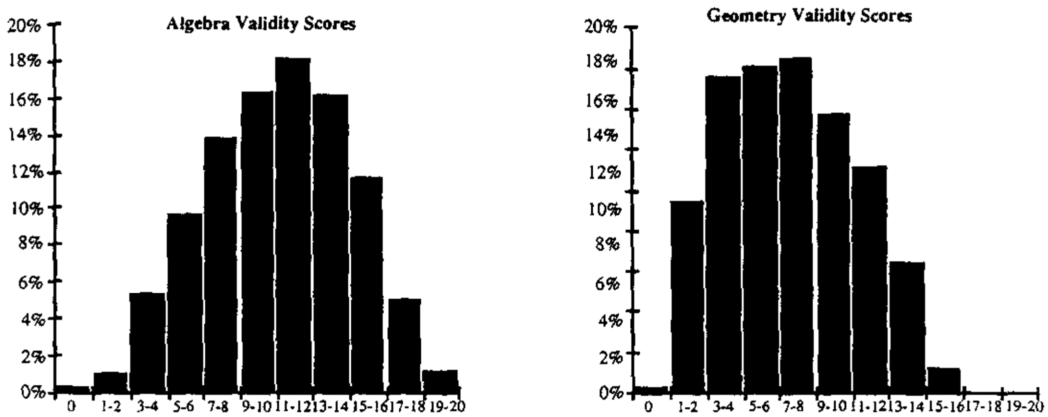


Figure 24: Distribution of validity scores.

These histograms show that in algebra, the modal group is 11-12, the distribution of AVS scores is symmetrical with a range of 0-20, and a small group of students obtained perfect scores. In geometry, the modal group is 7-8, the distribution is negatively-skewed with a range of 0-16, and not one student obtained a perfect score.

Taking into account all these findings, we report:

F47. Students are considerably better in algebra than in geometry at assessing whether an argument is correct and whether it is always or only sometimes true.

12. MULTILEVEL MODELLING: STUDENTS' VALIDITY SCORES

The next step in the analysis was to identify which, if any, of the Level 1 and Level 2 variables are associated with students' validity scores and whether the scores varied from school to school. Multilevel models of the two scores were constructed. Table 22 below presents the estimated effects of the significant variables in the final models.

	Algebra Validity Scores			Geometry Validity Scores		
Base group mean	9.29	(0.244)		5.39	(0.261)	
Variables						
Level 1						
Views of role of proof						
Truth	0.60	(0.159)	*0.073*	0.37	(0.151)	*0.041*
Student characteristics						
KS3 score	1.62	(0.133)	*0.284*	1.70	(0.130)	*0.234*
Responses to questionnaire						
Proof as general (algebra)	0.62	(0.157)	*0.074*	0.55	(0.147)	*0.061*
Proof as general (geometry)	0.66	(0.214)	*0.060*	0.75	(0.200)	*0.065*
Level 2						
School factors						
Selection procedures				0.90	(0.423)	*0.078*
Curriculum factors						
% GCSE higher tier	0.02	(0.004)	*0.145*	0.01	(0.005)	*0.076*
Approaches to teaching proof						
As a separate topic	0.81	(0.320)	*0.074*			
Notes: Standard errors shown in brackets; standardised effects shown in italics.						

Table 22: The estimated effects of the significant variables on the two validity scores.

Table 22 indicates that a similar set of variables is associated with student validity scores in algebra and in geometry, with the same four Level 1 variables significant for both scores. Once again we found that students' scores increase as Key Stage 3 test scores increase, and we note that this variable has a significant influence on all our output measures.

F48. The higher the Key Stage 3 test score, the better students are at evaluating arguments in terms of correctness and generality.

The other Level 1 variables concern how far students regarded proof as about establishing the truth of a statement and whether they had a sense of the generality of a valid mathematical proof. The latter point may not seem surprising, as deciding if a valid proof is true for a subset of cases might seem to point to a similar 'sense of proof' as being able to assess the correctness and scope of validity of a particular series of arguments. However, what is interesting is that this 'sense of proof' seems to be present in the assessment of very different types of argument and 'transfers' from geometry to algebra and vice versa.

F49. Students who regard proof as about establishing the truth of a statement are better at evaluating arguments in terms of their correctness and generality.

Level 2 variables also have significant effects and two associated with higher scores are of note: the % of students being entered for GCSE higher tier in both algebra and geometry; and teaching proof as a separate topic in algebra. We therefore report two findings:

F50. Students from classes with a large percentage of students expected to sit the higher-tier GCSE paper are likely to be better at evaluating arguments in terms of their correctness and generality than similar students from classes where more will be entered for the middle-tier paper.

F51. Students from classes where proof is taught as a separate topic are likely to be better at evaluating arguments in algebra in terms of their correctness and generality than similar students from classes where this does not happen.

SCHOOL DIFFERENCES

We now look at variation in the AVS and GVS scores of students from different schools to investigate if there were schools in which students obtained rather higher (or lower) scores than expected after adjustments according to the significant variables in the models have been made. Table 23 below presents the random effects associated with the variance component model for each validity score.

Random effects	Algebra Validity Score	Geometry Validity Score
Level 1 variation	12.2 (0.369)	10.64 (0.321)
Level 2 variation	0.83 (0.195)	1.89 (0.342)
Intra-school correlation	8.2%	15.1%
Notes: Standard errors in brackets.		

Table 23: Random effects for models of validity scores.

Table 23 shows that there was more variation in the performances of students *within* schools rather than *between* schools. Figure 25 presents the plots of the 95% uncertainty intervals around the school residuals for both algebra and geometry validity scores. They provide graphical illustration of the overlap between schools, but once again indicate that there was more variation between schools in geometry than in algebra.

F52. Student ability to assess the correctness and generality of an argument varies more in geometry than in algebra according to school attended.

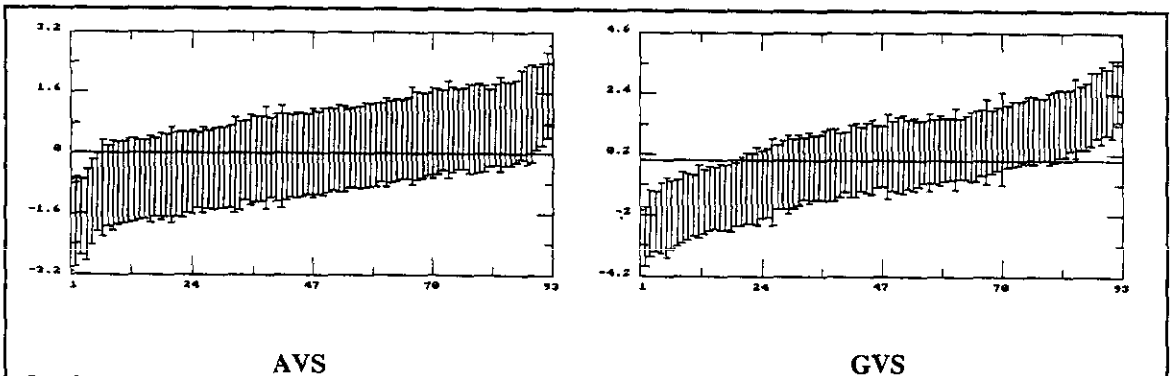


Figure 25: School deviations from predicted means: 95% uncertainty intervals around estimates for each school.

Random coefficient models were also fitted to each of the validity scores to explore whether or not the effects of explanatory variables are the same across all scores. No significant effects were found.

13. SUMMARY: STUDENTS' VALIDITY SCORES

When considering how well students are able to assess the generality of an argument, our analyses confirm a number of trends indicated in earlier sections: students obtain higher scores when constructing algebra proofs than geometry proofs, and they are also better at assessing the correctness and generality of the arguments presented in the algebra multiple-choice questions than those in geometry. As Key Stage 3 score increases, students can be expected to make more accurate assessments of arguments in both domains, as do students who are aware that proof involves verification. Students who are aware that valid proofs are general, are, not surprisingly, likely to do better at assessing the generality of the presented arguments than those who are not.

Again scores are associated with factors at the school-level and, in both algebra and geometry, students from classes where a large number of students are expected to sit the higher-tier GCSE paper are likely to be better at evaluating algebra and geometry arguments, than similar students from classes where more will be entered for the middle-tier paper.

Even after accounting for all the significant factors, students in some schools obtain higher scores than expected and, in others, students do less well than expected. This was particularly the case for validity scores in geometry, where our input variables accounted for less of the variation in scores that for algebra.

14. TEACHER AND SCHOOL EFFECTS

In the previous sections we have described how a range of individual, school, curriculum and teaching factors are associated with different student response patterns. While individual factors, and especially Key Stage 3 test score, tend to have the most consistent and pervasive influences across all the output measures, in general, some variation in student responses can be explained by looking to contextual factors at the school rather than the student level. Some of the contextual data we collected about the teachers and the schools, however, have no significant impact on student responses for any of the output measures we modelled. Of the school factors, the size of Year 10, the year in which students were first set for mathematics, and the number of sets into which they were placed turned out to be unrelated to student choices or scores.

F53 Neither the size of the Year 10 group nor the school policies for setting students for mathematics influence student responses for any of the output measures.

We also found that most of the teacher variables are not related to student performances. The teacher variables of sex, years of teaching experience and qualifications never improve the models of performance. Similarly, we found student choices to be completely independent of the choices selected by their teachers.

F54. Students' responses are not influenced by their teacher's sex, years of teaching experience or qualifications, nor by their teachers' responses to the questionnaire.

When examining school differences in the previous sections, we have described that, despite our centralised National Curriculum and even after taking into account all the contextual data we collected, some variation in student response at the school level was still evident. We have decided to conduct case-studies of a small number of schools to investigate the reason for these between-school differences. For all the output measures for which school differences were found, we selected the 10 schools in which student responses deviated most from the predicted patterns (i.e. the 5 schools with the highest estimates and the five with the lowest). School variation was found for all 6 scores and we examined our sample to see if the same schools were performing particularly well or badly for all scores or if different schools did better according to the different scores. On the whole, different schools did well for different scores; for example, a school where scores were much lower than expected for one score was not necessarily in the bottom five for another score. However, three classes appeared in the top 5 for at least half of the six scores, and one school appeared consistently amongst the bottom schools. We therefore rewrite finding F38:

F38. Although there is little between-school variation in students' scores, there are outlier schools whose students perform better or worse than predicted on 3 or more scores.

These schools are of particular interest, so we looked to see whether they were also performing in exceptional ways on the multiple-choice questions and in terms of the forms of argument used by students to construct their proofs. This was not generally found to be the case, so we have to rely on qualitative case study to seek out any distinguishing characteristics.

15. SUMMARY OF FINDINGS

A total of 54 findings have been reported in the previous sections. Below, these findings have been regrouped under the following headings: constructed proofs, comparisons of scores in algebra and geometry, generality of a valid proof, students' choices, influences on response patterns and unexplained variation in responses.

Constructed Proofs

F28. Students are unlikely to use deductive reasoning when constructing their own proofs. (p.42)

F32. **Empirical verification is the most popular form of argument used in constructing proofs, except in the case of an unfamiliar geometry proof.** (p.44)

F30. **Students construct narrative arguments more frequently than formal arguments.** (p.43)

F31. **Narrative arguments are more likely to be completely correct than other arguments, except when students attempt to prove an unfamiliar geometry statement.** (p.43)

Comparison of scores in algebra and geometry

F29. **Students are better at constructing proofs in algebra than in geometry.** (p.42)

F47. **Students are considerably better in algebra than in geometry at assessing whether an argument is correct and whether it is always or only sometimes true.** (p.60)

Generality of a valid proof

F26. **The majority of students know that once a statement has been proved it holds for all cases within its domain of validity.** (p.41)

Student choices

F27. **Students are better at choosing correct mathematical proofs than constructing them in both algebra and geometry.** (p.41)

F33. **In algebra but not geometry, students are more likely to construct empirical arguments than to choose them.** (p.44)

F11. **An argument whose generality is correctly appreciated is more likely to be chosen as a student's own approach than one that is not.** (p.23)

F12. **An argument felt to convince or explain is more likely to be selected as a student's own approach than one that is not, and the likelihood increases still further if it does both.** (p.30)

F2. **The form of argument selected to refute a proof or conjecture is influenced by whether the choice is for the student's own approach or for the best mark.** (p.21)

F3. **Students believe that proving or refuting a conjecture by means of a formally-presented analytic argument will receive the best mark.** (p.22)

F4. **Students are significantly more likely to select empirical arguments for their own approach than to receive the best mark.** (p.20)

F8. **Students are very unlikely to choose a simple counter-example for best mark, although they are significantly more likely to do so for their own approach.** (p.21)

F6. **An empirical verification of an algebra proof is very unlikely to be chosen to receive the best mark.** (p.20)

F5. **In algebra proofs, the most popular choice of presentation for a student's own approach has a prose-form, while the least popular is symbolic.** (p.20)

F6. **An empirical verification of an algebra proof is very unlikely to be chosen to receive the best mark.** (p.20)

F7. **A formal presentation of a proof is a more popular choice for a student's own approach in geometry than in algebra.** (p.20)

F9. **Students are more likely to choose a correct argument for best mark than for their own approach.** (p.22)

F10. **Students are more likely to choose an argument for their own approach if they believe it will receive the best mark.** (p.29)

Influences on response patterns

a. General mathematical attainment

F34. **Students with higher Key Stage 3 test scores are better at constructing proofs than those with lower scores.** (p.46)

F40. **In schools with better than predicted performance in constructing unfamiliar geometry proofs, students with high Key Stage 3 scores do particularly well.** (p.50)

F41. **With increase in Key Stage 3 test score, students become more likely to construct an argument with some relevant information, provided this is not empirical.** (p.53)

F48. **The higher the Key Stage 3 test score, the better students are at evaluating arguments in terms of correctness and generality.** (p.61)

F13. **Key Stage 3 test score influences the choice of argument for a student's own approach; as this score increases so does the student's preference for an argument which is not empirical.** (p.29)

F22. **As students' Key Stage 3 test score increases, so do their preferences for correct refutations, but also their preferences for incorrect formally-presented arguments.** (p.38) *b. Views of proof*

F1. Students are most likely to describe proof as about establishing the truth of a mathematical statement, although a substantial minority ascribe it an explanatory function and a further large number have little or no idea of the meaning of proof and what it is for. (p.18)

F36. Students who recognise the generality of a valid proof and appreciate its role in establishing the truth of a statement are better at constructing proofs than those who do not. (p.47)

F44. When attempting to construct an unfamiliar algebra proof, students who believe that proof has an explanatory function are less likely than those who do not, to construct formal arguments and more likely to construct narrative ones. (p.54)

F49. Students who regard proof as about establishing the truth of a statement are better at evaluating arguments in terms of their correctness and generality. (p.61)

F15. In most cases, student views of the role of proof influence their choice of argument for own approach and, in particular, students who have some idea of the role of proof are less likely to choose empirical arguments than those who do not. (p.30)

c. Student sex

F35. In algebra, girls construct better proofs than boys. (p.46)

F39. Schools make a significant difference to how well boys construct familiar algebra proofs, although this is not the case for girls. (p.49)

F42. Boys and girls use different forms of argument in algebra. (p.51)

F14. In most cases, girls and boys choose different arguments as their own approach to prove a statement. (p.26)

F23. There are differences in the choices of argument in response to a false conjecture between girls and boys. (p.38)

d. School, curriculum, teaching and teacher factors

F54. Students' responses are not influenced by their teacher's sex, years of teaching experience or qualifications, nor by their teachers' responses to the questionnaire. (p.64)

F53 Neither the size of the Year 10 group nor the school policies for setting students for mathematics influence student responses for any of the output measures. (p.63)

F16. Curriculum factors influence the arguments chosen for algebra proofs, with the main textbook and the hours of mathematics teaching each week exhibiting the most consistent effects. In particular, increasing the number of hours of mathematics teaching each week reduces the likelihood of students choosing an empirical argument. (p.30)

F17. Choices of proof in geometry are predominantly associated with student rather than school, curriculum and teacher factors. (p.26)

F21. Significant influences on students' responses to a false conjecture include student, school, curriculum and teaching variables. (p.36)

F24. Level 2 variables have significant influences on student responses to false conjectures, but only two have consistent effects: in geometry, students in selected schools are more likely than others to choose a correct refutation, but also more likely to choose an incorrect formally-presented argument; and in algebra, students following the London examination syllabus are more likely than others to choose the incorrect empirical verification. (p.38)

F43. The form of argument used in constructing a proof is associated with at least one Level 2 variable, with the main textbook used having the most consistent effect. (p.51)

F37. At least one curriculum or teaching factor is associated with students' constructed proof scores: variables of significance include the percentage of students in the class entering the GCSE higher-tier examination, the textbook followed and the expectation that students will write geometry proofs. (p.47)

F50. Students from classes with a large percentage of students expected to sit the higher-tier GCSE paper are likely to be better at evaluating arguments in terms of their correctness and generality than similar students from classes where more will be entered for the middle-tier paper. (p.62)

F45. Attending a school in which students are expected to write geometry proofs increases the likelihood that students will construct a formal argument when tackling an unfamiliar geometry proof. (p.54)

F51. Students from classes where proof is taught as a separate topic are likely to be better at evaluating arguments in algebra in terms of their correctness and generality than similar students from classes where this does not happen. (p.62)

Unexplained variation in responses

F38. Although there is little between-school variation in students' scores, there are outlier schools whose students perform better or worse than predicted on 3 or more scores. (p.64)

F52. Student ability to assess the correctness and generality of an argument varies more in geometry than in algebra according to school attended. (p.62)

F46. In general, the forms of argument used when students attempt to construct a mathematical proof differ according to school attended. (p.58)

F18. Although there is little between-school variation in the arguments selected as proofs of a conjecture, there are some schools where the students' preferences for analytical arguments are greater than predicted. (p.32)

F19. Student's choice for own approach varies more in geometry than in algebra according to school attended. (p.34)

F20. In geometry, the school attended can enhance students' preferences for a formal argument, regardless of whether or not it is correct. (p.35)

F25. There is no between-school variation in the arguments chosen in response to a false conjecture, after account is taken of all the variables that influence student choices. (p.39)

Variable codings from the School Questionnaire

Names	Code	Description
School level		
Sch_id	1-125	School identifier
schtype	LEA=1, Grant maintained=2, Roman Catholic=3, CofE=4	School type
schselec	Non=1, Some=2, Full=3	School academic selection
schsex	Girls=1, Boys=2, Mixed=3	School gender
area	Urban=1, Rural=2, Suburban=3	School location
no-y10	0-	No. of year 10 students
mathy10	Yes=1, No=2, Some=3	Whether the year 10 students are set for maths
wheny10	Y7=1, Y8=2, Y9=3, y10=4, n/a=5	The year when the year 10 students were first set for maths.
hoursy10	time in hours	Hours of maths. per week in Y10
Exam	Categories as appropriate	maths. curriculum exam syllabus
Main	Categories as appropriate	maths. curriculum main textbook
mathjust	Over=1, Under=2, Right=3	Maths. justification in the National Curriculum from the teacher
fmproof	Over=1, Under=2, Right=3	Description of formal proof in the National Curriculum from the teacher
State1H	1-2	High-achieving students: yes=1, no=2
State1O	3-4	Other students: yes=3, no=4
State2H	1-2	As State1H
State2O	3-4	As State1O
State3H	1-2	As State1H
State3O	3-4	As State1O
State4H	1-2	As State1H
State4O	3-4	As State1O
State5H	1-2	As State1H
State5O	3-4	As State1O
State6H	1-2	As State1H
State6O	3-4	As State1O
State7H	1-2	As State1H
State7O	3-4	As State1O
State8H	1-2	As State1H
State8O	3-4	As State1O
State9H	1-2	As State1H
State9O	3-4	As State1O
State10H	1-2	As State1H
State10O	3-4	As State1O

State11H	1-2	As State1H
State11O	3-4	As State1O
State12H	1-2	As State1H
State12O	3-4	As State1O
State13H	1-2	As State1H
State13O	3-4	As State1O
Class level		
Class_id	No.	Class identifier
Set	1= 1, 2 = 2, n/a = 0	Set level of class
SetsY10	No.	No. of set levels in Y10
GCSE%	Percentage	Approx. % of students to be given GCSE higher level paper in the class
tchrsex	F=1, M=2	Teacher's gender
tchyear	years	Teaching experience in year
Quali	Good = 1, Acceptable = 2, Weak = 3, Nil = 4, Various = 5	Level of qualification as specified in DFE teaching staff surveys (see statistical bulletin/Cockcroft)
tchA1	B=1, D=2, A=5, E=6, C=7	Teacher's own choice on question A1
tchA1tch	The same	Teacher thinks pupil thinks teacher's choice
tchA5	The same	As tchA1
tchA5tch	The same	As tchA1tch
tchA6	The same	As tchA1
tchA6tch	The same	As tchA1tch
tchG1	The same	As tchA1
tchG1tch	The same	As tchA1tch
tchG5	The same	As tchA1
tchG5tch	The same	As tchA1
tchG6	The same	As tchA1tch
tchA6tch	The same	As tchA1

Variable codings from the School Questionnaire

Names	Code	Description
Student level		
Student id	Sch_id + 1-60	Student's identifier
KS3test (SAT)	6-10	Student's National Curriculum level as specified in Year 9 Key Stage 3 test result
Sex	M=0, F=1	Student's gender
Age	months	Age in month at survey (14-15 year olds)
Pre-0	No answer=1, answered=0	Response for 'What is Proof in Maths. for?' For Role-0=0, the followings are recorded
Pre -1	1/0	Answers relating to "truth" =1, not=0
Pre -2	1/0	Answers relating to explanations =1, not=0
Pre -3	1/0	Answers relating to evidence =1, not=0
Pre -4	1/0	Answers relating to communication =1, not=0
Pre -5	1/0	Answers relating to discovery =1, not=0
Pre -6	1/0	Answers relating to "ability" =1, not=0
Pre -7	1/0	Answers relating to completeness =1, not=0
Pre -8	1/0	Answers relating to logic =1, not=0
Pre -9	1/0	Other answers =1, not=0
A1(std)	B=1, D=2, A=5, E=6, C=7	Student's choice for self on question A1
A1(tch)	As above	Student choice for best mark from the teacher on question A1
A1-A1	agree=1, don't know=2, disagree=3	A *mistake* in Arthur's answer
A1-A2	As above	*Always true* in Arthur's answer
A1-A3	As above	*Only-some* in Arthur's answer
A1-A4	As above	*Why* in Arthur's answer
A1-A5	As above	*Explain* in Arthur's answer
A1-B1	agree=1, don't know=2, disagree=3	*A mistake* in Bonnie's answer
A1-B2	As above	*Always true* in Bonnie's answer
A1-B3	As above	*Only-some* in Bonnie's answer
A1-B4	As above	*Why* in Bonnie's answer
A1-B5	As above	*Explain* in Bonnie's answer
A1-C1	agree=1, don't know=2, disagree=3	*A mistake* in Ceri's answer
A1-C2	As above	*Always true* in Ceri's answer
A1-C3	As above	*Only-some* in Ceri's answer
A1-C4	As above	*Why* in Ceri's answer
A1-C5	As above	*Explain* in Ceri's answer
A1-D1	agree=1, don't know=2, disagree=3	A *mistake* in Duncan's answer
A1-D2	As above	*Always true* in Duncan's answer
A1-D3	As above	*Only-some* in Duncan's answer
A1-D4	As above	*Why* in Duncan's answer
A1-D5	As above	*Explain* in Duncan's answer
A1-E1	agree=1, don't know=2, disagree=3	A *mistake* in Eric's answer

A1-E2	As above	*Always true* in Eric's answer
A1-E3	As above	*Only-some* in Eric's answer
A1-E4	As above	*Why* in Eric's answer
A1-E5	As above	*Explain* in Eric's answer
A2	A=1, B=2	Answer to question A2 on whether a proof is general (A) or specific (B).
A3-Y(std)	Yes=1, No=0	Does student prefer Yvonne's answer as choice for self
A3-Y(tch)	Yes=1, No=0	Does student prefer Yvonne's answer as choice for best mark from the teacher
A3-Y1	agree=1, don't know=2, disagree=3	A *mistake* in Yvonne's answer
A3-Y2	The same	*Always true* in Yvonne's answer
A3-Y3	The same	*Only-some* in Yvonne's answer
A3-Y4	The same	*Why* in Yvonne's answer
A3-Y5	The same	Explain in Yvonne's answer
A4-0	Not answered=1, answered=0	Form of presentation of student's constructed proof for question A4. For A4-0=0, the followings are recorded
A4-1	2 if main form 1 if included 0 if not included	2/1=Empirical by examples, 0=not
A4-2	As above	2/1=Exhaustive, 0=not
A4-3	As above	2/1=Empirical enactive, 0=not
A4-4	As above	2/1=Naive, 0=not
A4-5	As above	2/1=Formal correct, 0=not
A4-6	As above	2/1=Formal incorrect, 0=not
A4-7	As above	2/1=Narrative, 0=not
A4-8	As above	2/1=Visual, 0=not
A4-9	As above	2/1=Counter-example, 0=not
A4OVA	If A4-0 = 1,-1,-2 or if A4-4 = 0, then 0 else N=1, M=2, C=3	First scalar for student's own proof. 0 = no answer or naive answer 1 = No deductions, some basis for proof 2 = Partial proof 3 = Complete proof
A5(std)	H=1, J=5, F=6, G=7, I=9	Student's choice for self on question A5
A5(tch)	As above	Student choice for best mark from the teacher on question A5
A6(std)	L=1, N=5, M=6, K=7	Student's choice for self on question A6
A6(tch)	As above	Student choice for best mark from the teacher on question A6
A6-K1	agree=1, don't know=2, disagree=3	A *mistake* in Kate's answer
A6-K2	As above	*Always true* in Kate's answer
A6-K3	As above	*Only-some* in Kate's answer
A6-K4	As above	*Why* in Kate's answer
A6-K5	As above	*Explain* in Kate's answer
A6-L1	agree=1, don't know=2, disagree=3	A *mistake* in Leon's answer

A6-L2	As above	*Always true* in Leon's answer
A6-L3	As above	*Only-some* in Leon's answer
A6-L4	As above	*Why* in Leon's answer
A6-L5	As above	*Explain* in Leon's answer
A6-M1	agree=1, don't know=2, disagree=3	A *mistake* in Maria's answer
A6-M2	As above	*Always true* in Maria's answer
A6-M3	As above	*Only-some* in Maria's answer
A6-M4	As above	*Why* in Maria's answer
A6-M5	As above	*Explain* in Maria's answer
A6-N1	agree=1, don't know=2, disagree=3	A *mistake* in Nisha's answer
A6-N2	As above	*Always true* in Nisha's answer
A6-N3	As above	*Only-some* in Nisha's answer
A6-N4	As above	*Why* in Nisha's answer
A6-N5	As above	*Explain* in Nisha's answer
A7-0	Not answered=1, answered=0	Form of presentation of student's constructed proof for question A7. For A7-0=0, the followings are recorded
A7-1	2 if main form 1 if included 0 if not included	2/1=Empirical by examples, 0=not
A7-2	As above	2/1=Exhaustive, 0=not
A7-3	As above	2/1=Empirical enactive, 0=not
A7-4	As above	2/1=Naive, 0=not
A7-5	As above	2/1=Formal correct, 0=not
A7-6	As above	2/1=Formal incorrect, 0=not
A7-7	As above	2/1=Narrative, 0=not
A7-8	As above	2/1=Visual, 0=not
A7-9	As above	2/1=Counter-example, 0=not
A7OVA	If A7-0 = 1,-1,-2 or if A7-4 = 0, then 0 else N=1, M=2, C=3	First scalar for student's own proof. 0 =, some basis for proof 2 = Partial no answer or naive answer 1 = No deductions proof 3 = Complete proof
G1(std)	D=1, A=3, C=5, B=6, E=7	Student's choice for self on question G1
G1(tch)	As above	Student choice for best mark from the teacher on question G1
G1-A1	agree=1, don't know=2, disagree=3	A *mistake* in Amanda's answer
G1-A2	As above	*Always true* in Amanda's answer
G1-A3	As above	*Only-some* in Amanda's answer
G1-A4	As above	*Why* in Amanda's answer
G1-A5	As above	*Explain* in Amanda's answer
G1-B1	agree=1, don't know=2, disagree=3	*A mistake* in Barry's answer
G1-B2	As above	*Always true* in Barry's answer
G1-B3	As above	*Only-some* in Barry's answer
G1-B4	As above	*Why* in Barry's answer

G1-B5	As above	*Explain* in Barry's answer
G1-C1	agree=1, don't know=2, disagree=3	*A mistake* in Cynthia's answer
G1-C2	As above	*Always true* in Cynthia's answer
G1-C3	As above	*Only-some* in Cynthia's answer
G1-C4	As above	*Why* in Cynthia's answer
G1-C5	As above	*Explain* in Cynthia's answer
G1-D1	agree=1, don't know=2, disagree=3	A *mistake* in Dylan's answer
G1-D2	As above	*Always true* in Dylan's answer
G1-D3	As above	*Only-some* in Dylan's answer
G1-D4	As above	*Why* in Dylan's answer
G1-D5	As above	*Explain* in Dylan's answer
G1-E1	agree=1, don't know=2, disagree=3	A *mistake* in Ewan's answer
G1-E2	As above	*Always true* in Ewan's answer
G1-E3	As above	*Only-some* in Ewan's answer
G1-E4	As above	*Why* in Ewan's answer
G1-E5	As above	*Explain* in Ewan's answer
G2	A=1, B=2	Answer to question G2 on whether a proof is general (A) or specific (B).
G3-Y(std)	Yes=1, No=0	Does student prefer Yorath's answer as choice for self
G3-Y(tch)	Yes=1, No=0	Does student prefer Yorath's answer as choice for best mark from the teacher
G3-Y1	agree=1, don't know=2, disagree=3	A *mistake* in Yorath's answer
G3-Y2	As above	*Always true* in Yorath's answer
G3-Y3	As above	*Only-some* in Yorath's answer
G3-Y4	As above	*Why* in Yorath's answer
G3-Y5	As above	Explain in Yorath's answer

G4-0	Not answered=1, answered=0	Form of presentation of student's constructed proof for question G4. For G4-0=0, the followings are recorded
G4-1	2 if main form 1 if included 0 if not included	2/1=Empirical by examples, 0=not
G4-2	As above	2/1=Exhaustive, 0=not
G4-3	As above	2/1=Empirical enactive, 0=not
G4-4	As above	2/1=Naive, 0=not
G4-5	As above	2/1=Formal correct, 0=not
G4-6	As above	2/1=Formal incorrect, 0=not
G4-7	As above	2/1=Narrative, 0=not
G4-8	As above	2/1=Visual, 0=not
G4-9	As above	2/1=Counter-example, 0=not
G4OVA	If G4-0 = 1,-1,-2 or if G4-4 = 0, then 0 else N=1, M=2, C=3	First scalar for student's own proof. 0 = no answer or naive answer 1 = No deductions, some basis for proof 2 = Partial proof 3 = Complete proof
G5(std)	H=4, J=5, F=6, G=7, I=9	Student's choice for self on question G5
G5(tch)	As above	Student choice for best mark from the teacher on question G5
G6(std)	K=1, L=5, N=6, M=7	Student's choice for self on question G6
G6(tch)	K=1, L=5, N=6, M=7	Student choice for best mark from the teacher on question G6
G6-K1	agree=1, don't know=2, disagree=3	A *mistake* in Kobi's answer
G6-K2	As above	*Always true* in Kobi's answer
G6-K3	As above	*Only-some* in Kobi's answer
G6-K4	As above	*Why* in Kobi's answer
G6-K5	As above	*Explain* in Kobi's answer
G6-L1	agree=1, don't know=2, disagree=3	A *mistake* in Linda's answer
G6-L2	As above	*Always true* in Linda's answer
G6-L3	As above	*Only-some* in Linda's answer
G6-L4	As above	*Why* in Linda's answer
G6-L5	As above	*Explain* in Linda's answer
G6-M1	agree=1, don't know=2, disagree=3	A *mistake* in Marty's answer
G6-M2	As above	*Always true* in Marty's answer
G6-M3	As above	*Only-some* in Marty's answer
G6-M4	As above	*Why* in Marty's answer
G6-M5	As above	*Explain* in Marty's answer
G6-N1	agree=1, don't know=2, disagree=3	A *mistake* in Natalie's answer
G6-N2	As above	*Always true* in Natalie's answer
G6-N3	As above	*Only-some* in Natalie's answer
G6-N4	As above	*Why* in Natalie's answer

G6-N5	As above	*Explain* in Natalie's answer
G7-0	Not answered=1, answered=0	Form of presentation of student's constructed proof for question G4. For G4-0=0, the followings are recorded
G7-1	2 if main form 1 if included 0 if not included	2/1=Empirical by examples, 0=not
G7-2	As above	2/1=Exhaustive, 0=not
G7-3	As above	2/1=Empirical enactive, 0=not
G7-4	As above	2/1=Naive, 0=not
G7-5	As above	2/1=Formal correct, 0=not
G7-6	As above	2/1=Formal incorrect, 0=not
G7-7	As above	2/1=Narrative, 0=not
G7-8	As above	2/1=Visual, 0=not
G7-9	As above	2/1=Counter-example, 0=not
G7OVA	If G4-0 = 1,-1,-2 or if G4-4 = 0, then 0 else N=1, M=2, C=3	First scalar for student's own proof. 0 = no answer or naive answer 1 = No deductions, some basis for proof 2 = Partial proof 3 = Complete proof

List of derived variables referred to in technical report

Role of Proof		
None	1/0	1 if Role-0 = 1
Truth	1/0	1 if Role-1, Role-7, and/or Role-8 = 1
Social	1/0	1 if Role-2, Role-3 and/or Role-4 = 1
Discovery	1/0	1 if Role-5 = 1
Ability	1/0	1 if Role-6 = 1
A1(std)	1 - 8	Student's choice for self on question A1 or A3 if they would change their mind.
A1(tch)	As above	Student's choice for best mark from teacher on question A1 or A3 if they would change their mind.
Validity ratings (Algebra)		
vA1-A	2=correct validity evaluations 1=partially correct validity evaluations 0=incorrect validity evaluations	Validity rating for A1, Arthur's answer Derived from first three evaluations *mistake* *always true* *sometimes true*
vA1-B	As above	Validity rating for A1, Bonnie's answer
vA1-C	As above	Validity rating for A1, Ceri's answer
vA1-D	As above	Validity rating for A1, Duncan's answer
vA1-E	As above	Validity rating for A1, Eric's answer
vA1-Y	As above	Validity rating for A1, Yvonne's answer
vA6-K	As above	Validity rating for A6, Kate's answer
vA6-L	As above	Validity rating for A6, Leon's answer
vA6-M	As above	Validity rating for A6, Maria's answer

vA6-N	As above	Validity rating for A6, Nisha's answer
EvalAl	0-20	Sum of total validity rating for all algebra answers.
G1(std)	D=1, A=3, C=5, B=6, E=7 or if G3-Y(std) = 1, 8	Student's choice for self on question G1 or G3 if they would change their mind.
G1(tch)	As above	Student's choice for best mark from teacher on question G1 or G3 if they would change their mind.
Validity ratings (Geometry)		
vG1-A	2=correct validity evaluations 1=partially correct validity evaluations 0=incorrect validity evaluations	Validity rating for G1, Amanda's answer Derived from first three evaluations *mistake* *always true* *sometimes true*
vG1-B	As above	Validity rating for G1, Barry's answer
vG1-C	As above	Validity rating for G1, Cynthia's answer
vG1-D	As above	Validity rating for G1, Dylan's answer
vG1-E	As above	Validity rating for G1, Ewan's answer
vG1-Y	As above	Validity rating for G1, Yorath's answer
vG6-K	As above	Validity rating for G6, Kobi's answer
vG6-L	As above	Validity rating for G6, Linda's answer
vG6-M	As above	Validity rating for G6, Marty's answer
vG6-N	As above	Validity rating for G6, Natalie's answer
EvalGeo	0-20	Sum of total validity rating for all geometry answers
Students Constructed Proofs		
MAIN A4	0 - 8	Main form of presentation of student's constructed proof for question A4
MAIN A7	0 - 9	Main form of presentation of student's constructed proof for question A7
MAIN G4	0 - 8	Main form of presentation of student's constructed proof for question G4
MAIN G7	0 - 8	Main form of presentation of student's constructed proof for question G7
CPAlg	**0-6**	**Constructed proof (0, 1,2, or 3) score for A4 and A7 combined.** **(A4OVA +A7OVA)**
CPGeo	**0-6**	: **Constructed proof (0, 1,2, or 3) score for G4 and G7 combined.** **(G4OVA +G7OVA)**
Explanatory Ratings		
eA1-A	2=explains public and private 1=explains public or private 0=doesn't explain	Explanatory rating for A1, Arthur's answer Derived from last two evaluations *shows you why* *explains to someone in your class*
eA1-B	As above	Explanatory rating for A1, Bonnie's answer

eA1-C	As above	Explanatory rating for A1, Ceri's answer
eA1-D	As above	Explanatory rating for A1, Duncan's answer
eA1-E	As above	Explanatory rating for A1, Eric's answer
eA1-Y	As above	Explanatory rating for A1, Yvonne's answer
eA6-K	As above	Explanatory rating for A6, Kate's answer
eA6-L	As above	Explanatory rating for A6, Leon's answer
eA6-M	As above	Explanatory rating for A6, Maria's answer
eA6-N	As above	Explanatory rating for A6, Nisha's answer
eG1-A	As above	Explanatory rating for G1, Amanda's answer
eG1-B	As above	Explanatory rating for G1, Amanda's answer
eG1-C	As above	Explanatory rating for G1, Barry's answer
eG1-D	As above	Explanatory rating for G1, Cynthia's answer
eG1-E	As above	Explanatory rating for G1, Dylan's answer
eG1-Y	As above	Explanatory rating for G1, Ewan's answer
eG6-K	As above	Explanatory rating for G6, Kobi's answer
eG6-L	As above	Explanatory rating for G6, Linda's answer
eG6-M	As above	Explanatory rating for G6, Marty's answer
eG6-N	As above	Explanatory rating for G6, Natalie's answer

Note

The code for the missing or non-response without information can be -1. That for the informative non-response (such as 'not enough time', 'never heard about') can be -2 where it is necessary.

INSTITUTE OF
EDUCATION
UNIVERSITY OF LONDON

Appendix 1

The Proof Questionnaire

Justifying and Proving in School Mathematics

Funded by the Economic and Social Research Council

You are going to complete a survey that is all about proof.

Before you start, write below everything you know about proof in mathematics and what it is for.

Please do not write in this space

A

B

C

D

E

F

G

H

I

Algebra

A1. Arthur, Bonnie, Ceri, Duncan and Eric were trying to prove whether the following statement is true or false:

When you add any 2 even numbers, your answer is always even.

Please do not write in this space

> *Arthur's answer*
>
> a is any whole number
>
> b is any whole number
>
> $2a$ and $2b$ are any two even numbers
>
> $2a + 2b = 2(a + b)$
>
> *So Arthur says it's true.*

> *Bonnie's answer*
>
> $2 + 2 = 4$ $\quad$ $4 + 2 = 6$
>
> $2 + 4 = 6$ $\quad$ $4 + 4 = 8$
>
> $2 + 6 = 8$ $\quad$ $4 + 6 = 10$
>
> *So Bonnie says it's true.*

> *Ceri's answer*
>
> Even numbers are numbers that can be divided by 2. When you add numbers with a common factor, 2 in this case, the answer will have the same common factor.
>
> *So Ceri says it's true.*

> *Duncan's answer*
>
> Even numbers end in 0 2 4 6 or 8. When you add any two of these the answer will still end in 0 2 4 6 or 8.
>
> *So Duncan says it's true.*

> *Eric's answer*
>
> Let x = any whole number, y = any whole number
>
> $x + y = z$
>
> $z - x = y$
>
> $z - y = x$
>
> $z + z - (x + y) = x + y = 2z$
>
> *So Eric says it's true.*

From the above answers, choose **one** which would be closest to what you would do if you were asked to answer this question. [ ]

1B
2D
5A
6E
7C

For each of the following, circle whether you agree, don't know or disagree.

Please do not write in this space

The statement is:

When you add any 2 even numbers, your answer is always even.

	agree	don't know	disagree
Arthur's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** even numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Bonnie's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** even numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Ceri's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** even numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Duncan's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** even numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Eric's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** even numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

A2.

Please do not write in this space

Suppose it has now been proved that:

When you add any 2 even numbers, your answer is always even.

Zach asks what needs to be done to prove whether:

When you add 2 even numbers that are square, your answer is always even.

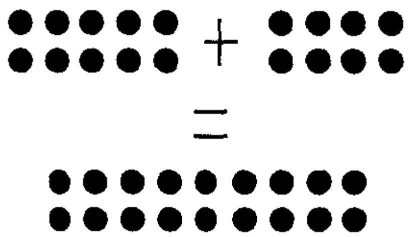
Tick either A or B.

(A) Zach doesn't need to do anything, the first statement has already proved this. ☐ A

(B) Zach needs to construct a new proof. ☐ B

A3. Yvonne drew the following picture for her answer to question A1:

Yvonne's answer



So Yvonne says it's true.

Please do not write in this space

Would you choose Yvonne's answer instead of your previous choice as the one closest to what you would do?

yes ☐ no ☐ Y N

Would you choose Yvonne's answer instead of your previous choice as the one your teacher would give the best mark?

yes ☐ no ☐ Y N

For each of the following circle whether you agree, don't know or disagree.

Yvonne's answer:	agree	don't know	disagree
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** even numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

A4. Prove whether the following statement is true or false. Write down your answer in the way that would get you the best mark you can.

When you add any 2 odd numbers, your answer is always even.

My answer

Please do not write in this space

0
1
2
3
4
5
6
7
8
9

N
M
C

A5. Farhana, Gary, Hamble, Iris and Julie were trying to prove whether the following statement is true or false:

Please do not write in this space

When you add any 3 consecutive numbers, your answer is always even.

Farhana's answer

x is any whole number.

$x+(x+1)+(x+2)=3x+3$

$3+3=6$

6 is divisible by 2

So Farhana says it's true.

Gary's answer

If the first number is even, then the second must be odd and the third must be even. This combination will always add up to be odd.

So Gary says it's false.

Hamble's answer

$3+4+5=12$

$11+12+13=36$

$35+36+37=108$

$107+108+109=324$

So Hamble says it's true.

Iris's answer

$2+3+4=9$

So Iris says it's false.

Julie's answer

Suppose first number is even, say $2x$.

$2x+(2x+1)+(2x+2)=6x+3$

$6x$ is even

$\therefore 6x+3$ is odd

So Julie says it's false

From the above answers, choose **one** which would be closest to what you would do if you were asked to answer this question. [ ]

1H 5J 6F 7G 9I

From the above answers, choose the **one** to which your teacher would give the best mark. [ ]

1H 5J 6F 7G 9I

A6. Kate, Leon, Maria and Nisha were asked to prove whether the following statement is true or false:

Please do not write in this space

When you multiply any 3 consecutive numbers, your answer is always a multiple of 6.

Kate's answer

A multiple of 6 must have factors of 3 and 2.
If you have three consecutive numbers, one will be a multiple of 3 as every third number is in the three times table.
Also, at least one number will be even and all even numbers are multiples of 2.
If you multiply the three consecutive numbers together the answer must have at least one factor of 3 and one factor of 2.

So Kate says it's true.

Leon's answer

$1 \times 2 \times 3 = 6$

$2 \times 3 \times 4 = 24$

$4 \times 5 \times 6 = 120$

$6 \times 7 \times 8 = 336$

So Leon says it's true.

Maria's answer

x is any whole number

$$x \times (x+1) \times (x+2) = (x^2 + 2) \times (x+2)$$
$$= x^3 + x^2 + 2x^2 + 2x$$

Cancelling the x's gives $1 + 1 + 2 + 2 = 6$

So Maria say it's true.

Nisha's answer

Of the three consecutive numbers, the first number is either:
EVEN which can be written $2a$ (a is any whole number) or,
ODD which can be written $2b$ - 1 (b is any whole number).

If EVEN
$2a \times (2a+1) \times (2a+2)$ is a multiple of 2.
and either a is a multiple of 3 DONE
or a is not a multiple of 3
$\therefore$ $2a$ is not a multiple of 3
$\therefore$ Either $(2a+1)$ is a multiple of 3 or $(2a+2)$ is a multiple of 3 DONE

If ODD
$(2b-1) \times 2b \times (2b+1)$ is a multiple of 2
and either b is a multiple of 3 DONE
or b is not a multiple of 3
$\therefore$ 2b is not a multiple of 3
$\therefore$ Either $(2b-1)$ is a multiple of 3 or $(2b+1)$ is a multiple of 3 DONE

So Nisha says it's true.

From the above answers, choose **one** which would be closest to what you would do if you were asked to answer this question.

1L
5N
6M
7K

From the above answers, choose the **one** to which your teacher would give the best mark.

1L
5N
6M
7K

For each of the following, circle whether you agree, don't know or disagree.

Please do not write in this space

The statement is:

When you multiply any 3 consecutive numbers, your answer is always a multiple of 6.

	agree	don't know	disagree
Kate's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows the statement is true for **some** consecutive numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Leon's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows the statement is true for **some** consecutive numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Maria's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows the statement is true for **some** consecutive numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Nisha's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows the statement is true for **some** consecutive numbers	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

A7. Prove whether the following statement is true or false. Write your answer in a way that would get you the best mark you can.

Please do not write in this space

If p and q are any two odd numbers, $(p+q)\times(p-q)$ is always a multiple of 4.

My answer

0
1
2
3
4
5
6
7
8
9

N
M
C

Geometry

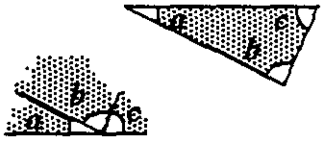
G1. Amanda, Barry Cynthia, Dylan, and Ewan were trying to prove whether the following statement is true or false:

Please do not write in this space

When you add the interior angles of any triangle, your answer is always 180°.

Amanda's answer

I tore the angles up and put them together.

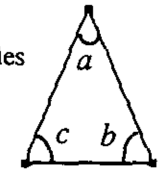


It came to a straight line which is 180°. I tried for an equilateral and an isosceles as well and the same thing happened.

So Amanda says it's true.

Barry's answer

I drew an isosceles triangle, with c equal to 65°.



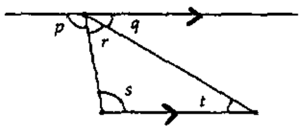
Statements	Reasons
$a = 180° - 2c$.......	Base angles in isosceles triangle equal
$a = 50°$..............	180° - 130°
$b = 65°$..............	180° - $(a + c)$
$c = b$.................	Base angles in isosceles triangle equal

$\therefore\ a + b + c = 180°$

So Barry says it's true.

Cynthia's answer

I drew a line parallel to the base of the triangle



Statements	Reasons
$p = s$..........................	Alternate angles between two parallel lines are equal
$q = t$..........................	Alternate angles between two parallel lines are equal
$p + q + r = 180°$........	Angles on a straight line

$\therefore\ s + t + r = 180°$

So Cynthia says it's true.

Dylan's answer

I measured the angles of all sorts of triangles accurately and made a table.

a	*b*	*c*	*total*
110	34	36	180
95	43	42	180
35	72	73	180
10	27	143	180

They all added up to 180°.

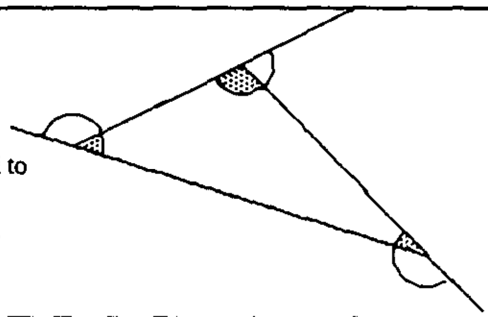
So Dylan says it's true.

Ewan's Answer

If you walk all the way around the edge of the triangle, you end up facing the way you began. You must have turned a total of 360°.

You can see that each exterior angle when added to the interior angle must give 180° because they make a straight line. This makes a total of 540°. 540° - 360° = 180°.

So Ewan says it's true.



From the above answers, choose **one** which would be closest to what you would do if you were asked to answer this question. [ ]

1D
3A
5C
6B
7E

From the above answers, choose the **one** to which your teacher would give the best mark. [ ]

1D
3A
5C
6B
7E

For each of the following, circle whether you agree, don't know or disagree.

Please do not write in this space

The statement is:

When you add the interior angles of any triangle, your answer is always 180°.

	agree	don't know	disagree
Amanda's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** triangles	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Barry's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** triangles	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Cynthia's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** triangles	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Dylan's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** triangles	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3
Ewan's answer:			
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** triangles	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

G2. Suppose it has now been proved that

Please do not write in this space

When you add the interior angles of any triangle, your answer is always 180°.

Zoe asks what needs to be done to prove whether:

When you add the interior angles of any right-angled triangle, your answer is always 180°.

Tick either A or B:

(A) Zoe doesn't need to do anything, the first statement has already proved this. ❑

A

(B) Zoe needs to construct a new proof. ❑

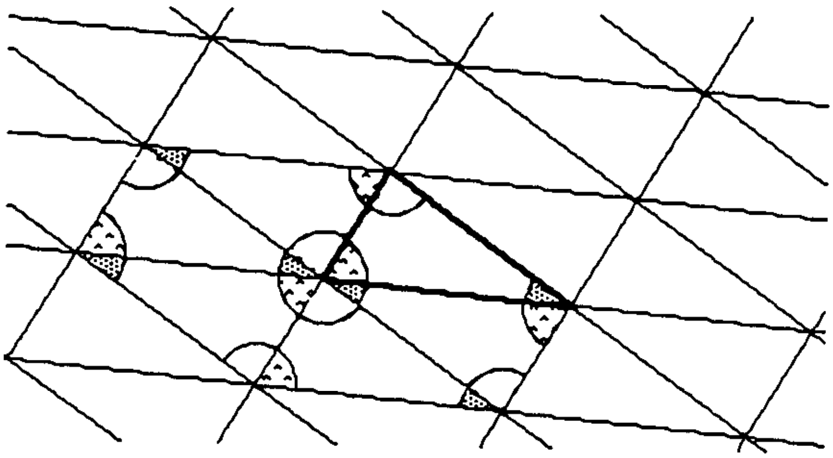
B

G3. Yorath gave the following answer to question G1:

Please do not write in this space

Yorath's answer

I drew a tessellation of triangles and marked all the equal angles.



I know that the angles round a point add up to 360°.

So Yorath says it's true.

Would you choose Yorath's answer instead of your previous choice as the one closest to what you would do ?

yes ☐ no ☐ Y N

Would you choose Yorath's answer instead of your previous choice as the one your teacher would give the best mark?

yes ☐ no ☐ Y N

For each of the following circle whether you agree, don't know or disagree

Yorath's answer:	agree	don't know	disagree
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** triangles	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

G4. Prove whether the following statement is true or false. Write your answer in a way that would get you the best mark you can.

If you add the interior angles of any quadrilateral, your answer is always 360°.

My answer

Please do not write in this space

0

1

2

3

4

5

6

7

8

9

N

M

C

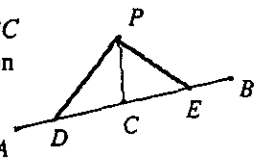
G5. Frank, Gail, Harriet, Irene and Jacob were trying to prove whether the following statement is true or false:

Please do not write in this space

The shortest distance between any point P and a line segment AB is the line joining P to C, where C is the midpoint of AB.

Frank's answer

E is any point on BC and D is any point on AC.

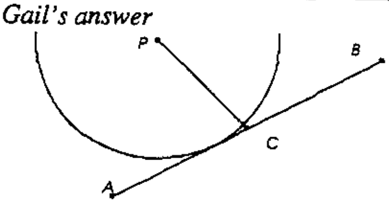


Statement	Reason
$AC = BC$	C is the midpoint
$CE^2 + PC^2 = PE^2$	Pythagoras theorem
$CD^2 + PC^2 = PD^2$	Pythagoras theorem
$PC \leq PE$	CE is greater than 0
$PC \leq PD$	CD is greater than 0

$\therefore$ PC is the shortest distance

So Frank says it's true.

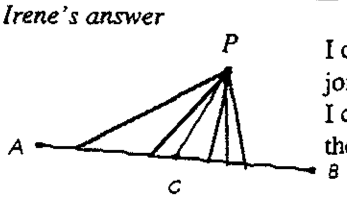
Gail's answer



I drew an arc with my compass using P as the centre and so that the arc just touched the line AB. The line from P to C crossed the circle showing that PC was not the shortest line.

So Gail says it's false.

Irene's answer



I drew some more lines joining P to AB
I can see that some of them are shorter than PC.

So Irene says it's false.

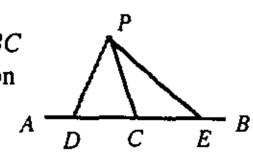
Harriet's answer

A straight line is always the shortest distance between two points.

So Harriet says it's true.

Jacob's answer

E is any point on BC and D is any point on AC.



Statement	Reason
If angle $PCE > 90°$	
angle $PEC < 90°$	Sum of angles in a triangle = 180°
and $PE > PC$	Longest side of triangle is opposite largest angle
But if angle $PCE > 90°$	
angle $PCD < 90°$	Sum of angles on a straight line = 180°
angle PDC can be > 90.......	Sum of angles in a triangle = 180°
PD can be < PC	Longest side of triangle is opposite largest angle

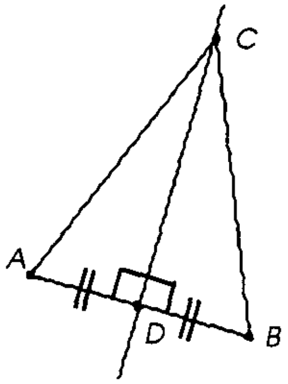
$\therefore$ PC is not always the shortest distance.

So Jacob says it's false.

From the above answers, choose **one** which would be closest to what you would do if you were asked to answer this question. 4H 5J 6F 7G 9I

From the above answers, choose the **one** to which your teacher would give the best mark. 4H 5J 6F 7G 9I

G6.

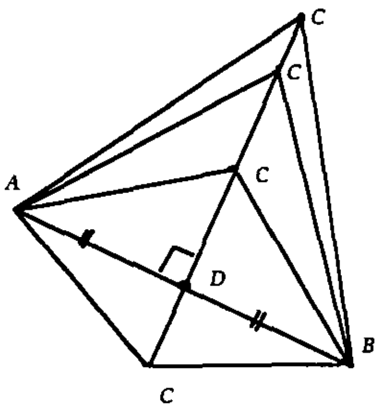


C is any point on the perpendicular bisector of AB. Kobi, Linda, Marty and Natalie were trying to prove whether the following statement is true or false:

Triangle ABC is always isosceles.

Please do not write in this space

Kobi's answer



I moved C to different places on the perpendicular bisector and measured AC and BC. They were always the same so the triangles were all isosceles.

So Kobi says it's true.

Linda's answer

Statement	Reason
$AD = BD$................	Bisector
$ADC = 90°$............	Perpendicular line
$BDC = 90°$............	Perpendicular line
$DC = DC$...............	Same line
$\Delta ADC = \Delta BDC$.....	Two sides and included angle the same.
$\therefore AC = BC$.	

So Linda says it's true.

Marty's answer

Because CD bisects AB at right angles, B is a reflection of A. So you could think of ABC as made up of two right angle triangles which are reflections of each other. This means the sides AC and BC will be the same length.

So Marty says it's true.

Natalie's answer

Statement	Reason
$ADC = 90°$........................	Perpendicular line
$BDC = 90°$........................	Perpendicular line
Angle CAB = angle CBD.	Base angles of an isosceles triangle equal
$\therefore AC = BC$.	

So Natalie says it's true.

From the above answers, choose **one** which would be closest to what you would do if you were asked to answer this question. [ ]

1K
5L
6N
7M

From the above answers, choose the **one** to which your teacher would give the best mark. [ ]

1K
5L
6N
7M

For each of the following circle whether you agree, don't know or disagree.

Please do not write in this space

The statement is:

Triangle ABC is always isosceles.

Kobi's answer:

	agree	don't know	disagree
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** positions of *C*	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

Linda's answer:

	agree	don't know	disagree
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** positions of *C*	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

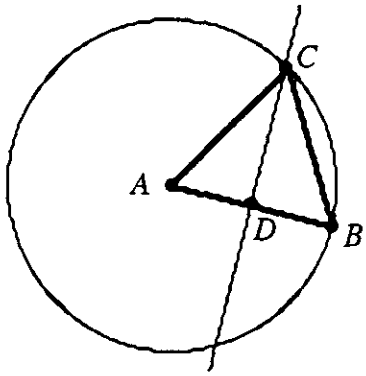
Marty's answer:

	agree	don't know	disagree
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** positions of *C*	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

Natalie's answer:

	agree	don't know	disagree
Has a **mistake** in it	1	2	3
Shows that the statement is **always true**	1	2	3
Only shows that the statement is true for **some** positions of *C*	1	2	3
Shows you **why** the statement is true	1	2	3
Is an easy way to **explain** to someone in your class who is unsure	1	2	3

G7.



A is the centre of a circle and AB is a radius. C is a point on the circumference where the perpendicular bisector of AB crosses the circle. Prove whether the following statement is true or false. Write your answer in a way that would get you the best mark you can.

Triangle *ABC* is always equilateral.

My answer

Please do not write in this space

0
1
2
3
4
5
6
7
8
9

N
M
C

INSTITUTE OF
EDUCATION
UNIVERSITY OF LONDON

Appendix 2

The School Questionnaire

Justifying and Proving in School Mathematics

Funded by the Economic and Social Research Council

SCHOOL QUESTIONNAIRE

School: ....

Town ....

County ....

School data
Please circle the number which best describes your school.

Type		Selection		Single-sex/mixed		Area	
County (LEA)	☐1	No academic selection	☐1	Girls-only	☐1	Urban	☐1
Grant maintained	☐2	Some academic selection	☐2	Boys-only	☐2	Rural	☐2
Roman Catholic	☐3	Full academic selection	☐3	Mixed-sex	☐3	Suburban	☐3
Anglican (CofE)	☐4						

Student data
Please attach a list of Year 9 SAT results for students who are completing the survey.

Year 10 data

Approx. no. of Y10 students ....

Are current Y10 students set for mathematics? Yes ☐1 No ☐2 For some lessons only ☐3

When were these students first set for mathematics? Y7 ☐1 Y8 ☐2 Y9 ☐3 Y10 ☐4 n/a ☐5

Please complete the following.

The students who are completing the survey come from set ☐ of ☐

Give approx % of students from this class who you predict will be entered for GCSE higher level paper ....

Mathematics curriculum data

Exam syllabus ....

Main textbook / Scheme ....

Hours of mathematics per week in Y10 ....

Proof in the curriculum
Please tick the statements which best describes your feelings about mathematical justification and formal proof in the National Curriculum.

Mathematical justification is	over-emphasised	☐1
	under-emphasised	☐2
	about right	☐3
Formal proof is	over-emphasised	☐1
	under-emphasised	☐2
	about right	☐3

Turn over

Please read though the following statements and circle those which most closely match your approach and that of your department.

	For high-achieving students		For other students	
The coverage of mathematical justification:				
is greater than the National Curriculum specifications	Yes 1	No 2	Yes 3	No 4
is much the same as the National Curriculum specifications	Yes 1	No 2	Yes 3	No 4
The coverage of formal proof:				
is greater than the National Curriculum specifications	Yes 1	No 2	Yes 3	No 4
is much the same as the National Curriculum specifications	Yes 1	No 2	Yes 3	No 4
Mathematical justification is addressed mainly through investigations	Yes 1	No 2	Yes 3	No 4
Mathematical justification is addressed as a topic area in its own right	Yes 1	No 2	Yes 3	No 4
Students are expected to show that mathematical statements are true	Yes 1	No 2	Yes 3	No 4
Students are expected to explain why mathematical statements are true	Yes 1	No 2	Yes 3	No 4
Students are expected to use some deduction	Yes 1	No 2	Yes 3	No 4
Students are expected to read formal geometric proofs	Yes 1	No 2	Yes 3	No 4
Students are expected to write formal geometric proofs	Yes 1	No 2	Yes 3	No 4
Students are expected to read formal algebraic proofs	Yes 1	No 2	Yes 3	No 4
Students are expected to write formal algebraic proofs	Yes 1	No 2	Yes 3	No 4

Any other comments:

Teacher data

Sex Female ☐ 1 Male ☐ 2

No. years teaching experience ....

Please specify type of qualification and subjects studied.

Type		Main subject (please specify)	Subsidiary subject (please specify)
Degree (not BEd)	☐ 1	....	....
BEd	☐ 2	....	....
PGCE	☐ 3	....	....
Cert Ed	☐ 4	....	....
Other	☐ 5	....	....

Would you be willing for us to approach you about making a second visit to interview a sample of the students who have completed the questionnaire? Yes ☐ 1 No ☐ 2

INSTITUTE OF
EDUCATION
UNIVERSITY OF LONDON

Appendix 3

Administration Procedures

Justifying and Proving in School Mathematics

Funded by the Economic and Social Research Council

Instructions for Administering Proof Survey

1. Altogether you will need **one hour and ten minutes** to complete the survey. Sometimes this may involve the students staying a little longer than the normal lesson time. If so, make sure you have negotiated this with the teacher before you go.

2. Please conduct the survey in a **formal** way.

3. Please do **not** help the students or intervene in their work.

4. Some students may be unfamiliar (or have forgotten) some of the **mathematical terms** in the survey. You will have been sent a large sheet with a list of mathematical definitions. Please **pin this up** in full view of all the students.

5. **Class Teacher**
   - The class teacher should **stay with you** during the completion of the survey.
   - **Proof questionnaire for the teacher**
     One of the questionnaires has a **red** cover. This is for the class teacher to fill in **while** their students are completing their questionnaires. Please give this to the teacher and ask them to:
     (i) Fill in what **they** would do for all questions "choose which would be closest to what you would do". (Questions A1, A5, A6, G1, G5 and G6).
     (ii) Fill in what **they** predict their students will write for all the questions "choose which your teacher would give the best mark". (Questions A1, A5, A6, G1, G5 and G6).
   - They will have been sent a **school questionnaire** asking for background information on the school and the students, but **please take a copy** with you in case this has been mislaid. **Make sure they have completed this before you leave.**
   - **Make sure you get the SAT results.**

6. You need to give a brief introduction (**5 minutes** max.):
   - They are taking part in a **nationwide** survey of year 10 students' ideas about proof.
   - The results are **very important** and relate to Attainment Target 1 in the Mathematics curriculum.
   - We are interested in **individual** views so they need to complete the survey on their own, but their identity will be kept **confidential**.
   - Many questions are structured in a similar way where there is a **mathematical statement** followed by a **number of answers** given by students who were trying to work out whether the statement was true or false (student proofs). Some answers may be right and some may be wrong, but there is **never only one right answer**. They will be asked a series of questions about these answers. They are given five questions about each student proof. On **each** question they need to circle whether they agree, don't know or disagree. They will also be asked to **construct an answer of their own** to prove a mathematical statement. They can base their answers on the examples, or come up with new answers. This is up to them.
   - They have **5** minutes to complete the white page, **30** minutes for the blue and **30** minutes for the pink. They **need to keep to time** as there is a lot to complete.
   - They are allowed to use **calculators** and other **mathematical aids** (rulers, compasses, etc.) if they wish.

7 After the introduction, you are ready to start.
   - Tell them they have **5** minutes to answer the question on what is proof (white page). When they have written as much as they want, they should **draw a line** under what they have written and move onto the next set of coloured sheets.
   - **After 5 minutes** announce that it is time to move on.
   - **After another 30 minutes** announce that they should move on to the next set of coloured sheets

8. If any student finishes early, then tell them **in this order** to:
   - Check through their answers.
   - Write what they **now** think proof in mathematics is for. They should add this on the white sheet **under** the line they drew.
   - If they still have time left over, give them a copy of the **red** sheet.

9. At the end of the time for the survey (65 minutes) tell the students to **stop working**, then tell them to go through their sheets and where they have left a question blank write in the space provided for the answer either **no response** if they could not do it, or **no time** if they had run out of time .

Check List

Have you collected all the proof questionnaires? ☐
Please check the number of scripts against class list

Do you have all the students' SAT results? ☐

Do you have the teacher's copy of the proof questionnaire (red cover) ☐

Please check that the correct sections have been completed ☐

Do you have the school questionnaire? ☐

Please check that **every** section of the school questionnaire is completed ☐

Comments

Add to the box below please any thoughts, observations or comments you would like to make about the administration of the questionnaire, etc.

Please send this sheet and all the data to

Lulu Healy
Mathematical Sciences
Institute of Education
University of London
20 Beford Way
London WC1H OAL

Tel: 0171 612 6678
Fax: 0171 612 6686

INSTITUTE OF
EDUCATION
UNIVERSITY OF LONDON

Appendix 4

Multilevel Models of Student Scores

Justifying and Proving in School Mathematics

Celia Hoyles, Lulu Healy
and Min Yang

Funded by the Economic and Social Research Council

In the following section, we briefly explain the models applied to analyse the scores associated with our dataset along with the assumptions underlying each of the models used.

A MULTILEVEL MODEL

A simple single level regression model can be written as:

$$y_i = \beta_0 + e_i \qquad (1)$$

where:

y_i = the score of the *i*th student

β_0 = the predicted score for the *i*th student

e_i = the departure of the *i*th student's actual score from the predicted score or the *residual.*

In the multilevel case, schools are also regarded as a random sample, hence (1) is re-expressed as:

$$y_{ij} = \beta_{0j} + e_{ij} \qquad (2)$$

where:

ij denotes the *i*th student in the *j*th school.

Unlike the conventional regression model, we assume that the regression parameter, β_0, is a random variable at the school level; namely the estimate may have a distribution around a population mean with a variance. Therefore at Level 2, we have the following model for predicting school means based on (2),

$$\beta_{0j} = \beta_0 + u_j \qquad (3)$$

where the term u_j refers to the residuals of school estimates from the overall mean estimate.

By substituting β_{0j} from (3) in (2), we have the complete model as:

$$y_{ij} = \beta_0 + u_j + e_{ij} \qquad (4)$$

where for the i^{th} student in the j^{th} school, u_j and e_{ij} are random variables (error terms) representing respectively, school-level deviations and additional student-level deviations from the expected value β_0 (the fixed effect). The u_j are assumed to be independently and normally distributed with a zero mean and a constant variance σ_u^2 to be estimated ($u_j \sim N(0,\sigma_u^2)$). The e_{ij} are also assumed to be independently normally distributed with zero mean and constant variation σ_e^2, also to be estimated ($e_{ij} \sim N(0,\sigma_e^2)$). These two error terms are further assumed not to be correlated, i.e., there is no covariance between the error terms from different levels.

The function of the error terms is to model any observed variation of the scores about the values predicted by the rest of the model. A non-zero estimate of σ_u^2 indicates that there is a component of the total variation which is associated with schools and that students' scores within a school are correlated. To examine the effects of other explanatory variables, we can simply add them into (4). In this study, we have variables at both Level 1 (student factors) and at Level 2 (the school, curriculum, teaching and teacher factors).

Two types of models were constructed, the variance component model for modelling students' proof scores, and the random coefficient model for modelling student choices. We start by describing the former simpler model.

THE VARIANCE COMPONENT MODEL

In the variance component model, we assume that there is only random variation around an estimate for β_0; that is, any estimates associated with the explanatory variables are the same for all schools. In this case, no further random terms are added.

So, the model of students' scores for their constructed proofs which is, for example, conditional on one Level 1 factor, sex, and one Level 2 factor, percentage of students entered for the GCSE higher tier, would be:

$$y_{ij} = \overbrace{\beta_0 + \beta_1 SEX_{ij} + \beta_2 GCSE_j}^{\text{Fixed Effects}} + \overbrace{u_j + e_{ij}}^{\text{Random Effects}} \qquad (5)$$

where for the *i*th student in the *j*th school:
y_{ij} is the raw score the student obtained for their constructed proof;
SEX_{ij} is a *Level 1* variable indicating the sex of the students, with value 0 for a male and 1 for a female.
$GCSE_j$ is a *Level 2* variable representing the percentage of the students in the class expected to be entered for the higher tier at GCSE, centred around the mean for all students in the sample (80%).
The model has two parts, fixed and random. In the fixed part of the model, the β's represent population parameters to be estimated. In our example, β_0 can be interpreted as the expected score for a male student from a class in which 80% of students are expected to be entered for the higher tier paper.
β_1 can be interpreted as the expected difference in scores associated with being a female and β_2 the expected difference in scores per percentage point of the GCSE variable.
Hence to obtain the predicted score of a female student in a class where 100% of students would take the higher tier GCSE paper, the following calculation would be made:
$\beta_0 + (\beta_1 \times 1) + (\beta_2 \times 20)$

In the random part of the variance component model, we have two error terms u_j and e_{ij} representing respectively Level 2 deviations and additional Level 1 deviations.

To generalise model (5) so as to allow general representations of the Level 1 and Level 2 variables, we use x_{pij} for Level 1 variables and x_{hj} for Level 2 variables. So, for example (5) above would be written using x_1 for SEX and x_2 for GCSE as follows:

$$y_{ij} = \overbrace{\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j}}^{\text{Fixed Effects}} + \overbrace{u_j + e_{ij}}^{\text{Random Effects}} \qquad (5)$$

With this notation, further explanatory variables can be added as desired to the model shown below in (6):

$$y_{ij} = \overbrace{\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + \beta_4 x_{4j}}^{\textit{Fixed Effects}} + \overbrace{u_j + e_{ij}}^{\textit{Random Effects}} \qquad (6)$$

where, for example:

$x_1 =$ a Level 1 dummy variable for SEX (0 for male, 1 for female)

$x_2 =$ a Level 1 variable for AGE, centred around the mean for all students in the sample.

$x_3 =$ a Level 2 variable representing the percentage of the class expected to be entered for the higher tier GCSE paper, centred around the mean for all students in the sample.

$x_4 =$ a Level 2 dummy variable indicating whether or not proof is taught as a topic in its own right.

Model (6) can be rewritten as:

$$y_{ij} = \overbrace{\beta_0 + \sum_{p=1}^{2} \beta_p x_{pij} + \sum_{h=3}^{4} \beta_h x_{hj}}^{\textit{Fixed Effects}} + \overbrace{u_j + e_{ij}}^{\textit{Random Effects}} \qquad (7)$$

Hence in general a model with n Level 1 variables and m Level 2 variables can be written as:

$$y_{ij} = \overbrace{\beta_0 + \sum_{p=1}^{n} \beta_p x_{pij} + \sum_{h=n+1}^{n+m} \beta_h x_{hj}}^{\textit{Fixed Effects}} + \overbrace{u_j + e_{ij}}^{\textit{Random Effects}} \qquad (8)$$

THE RANDOM COEFFICIENT MODEL

To model more complex variation, instead of assuming that the effect of a given explanatory variable will be the same in all schools, we allow for random variation in the fixed effects *across* schools. For example, we might find that the effects associated age are not the same in all schools. Let us consider the case for a model in which we include only this Level 1 variable, AGE. Substituting n = 1 and m= 0 into (8) we get a variance component model as follows:

$$y_{ij} = \overbrace{\beta_0 + \beta_1 x_{1ij}}^{\textit{Fixed Effects}} + \overbrace{u_j + e_{ij}}^{\textit{Random Effects}}$$

where

$x_l =$ a Level 1 variable for AGE, centred around the mean for all students in the sample

Recall (3) in which we replaced the β_0 in the simple regression model by β_{0j} to allow the average to vary across schools so that:

$$\beta_{0j} = \beta_0 + u_j \qquad (3)$$

We now adopt a similar approach for other parameters in the fixed part of the model. So in (9) we replace β_1 by β_{1j} where $\beta_{1j} = \beta_1 + v_j$ and v_j is a new random variable (error term) also representing school-Level deviations from average and again assumed to be normally distributed with a zero mean and a constant variance σ_v^2 to be estimated ($v_j \sim N(0, \sigma_v^2)$). Hence (9) becomes:

$$y_{ij} = \overbrace{\beta_0 + \beta_1 x_{1ij}}^{\textit{Fixed Effects}} + \overbrace{u_j + v_j x_{1ij} + e_{ij}}^{\textit{Random Effects}} \qquad (10)$$

We now have more than one variable at the same level which may be correlated so it is necessary to estimate the covariance between them. In total, three random parameters will be estimated at Level 2:

σ_u^2, the variance of the u_j
σ_v^2, the variance of the v_j
σ_{uv}, the covariance of the u_j with the v_j

The total variance at Level 2 is the sum of variances and covariances of these random variables and can be given as:

$$\text{var}(u_j + v_j x_{1ij}) = \sigma_u^2 + 2\sigma_{uv} x_{1ij} + \sigma_v^2 x_{1ij}^2 \qquad (11)$$

At Level 1 we still have only one random parameter to be estimated:
σ_e^2, the variance of the e_{ij}.

Up to this point, we have still been working with the assumption that Level 1 residuals have a constant variance, but this might not always be the case. Again using student age as our example, we might find that the variance of the student-level residuals increases (or decreases) as students get older.

Just as we can model variation at Level 2 by allowing the coefficient of the variable for AGE (x_1) to have a random component, such Level 1 variance can also be modelled by adding a further random term to the model. β_{1j} now becomes $\beta_1 + v_j + f_{ij}$ where f_{ij} is the additional student-level variation around the school means. The new model is given in (12):

$$y_{ij} = \overbrace{\beta_0 + \beta_1 x_{1ij}}^{\textit{Fixed Effects}} + \overbrace{u_j + v_j x_{1ij}}^{\textit{Random Effects (level 2)}} + \overbrace{e_{ij} + f_{ij} x_{1ij}}^{\textit{Random Effects (level 1)}} \qquad (12)$$

We still have the same fixed parameters, but now we have four random variables in the model, two at each level. Altogether this will lead to estimates for six random parameters: the variances of each of these error terms along with the covariances of the pairs at the same level. The three parameters at Level 2 will be as for model (10), and at Level 1 we also now have estimates for 3 random parameters:

σ_e^2 , the variance of the e_{ij}

σ_f^2, the variance of the f_{ij}

σ_{ef}, the covariance of the e_{ij} with the f_{ij}

The total variance at Level 1 can now be expressed as:

$$\text{var}(e_{ij} + f_{ij} x_{1ij}) = \sigma_e^2 + 2\sigma_{ef} x_{1ij} + \sigma_f^2 x_{1ij}^2 \qquad (13)$$

INSTITUTE OF
EDUCATION
UNIVERSITY OF LONDON

Appendix 5

Multilevel Models of Categorical Data

Justifying and Proving in School Mathematics

Min Yang, Lulu Healy
and Celia Hoyles

Funded by the Economic and Social Research Council

In this appendix, we briefly explain the models applied to analyse the categorical data associated with our dataset.

MULTINOMIAL MODELS FOR MULTIPLE CATEGORICAL OUTCOMES

Two types of questions from the survey produced categorical outcomes: the multiple choice questions and the form of argument students used in their constructed proofs. Multinomial models were constructed to obtain estimates for variables associated with these choices i.e. for each of the six multiple choice questions and the four constructed proof questionsand below we describe this kind of model using the first of the multiple choice question (A1, where student are asked to pick one from six possible choices as closest to how they would approach the task of proving that the sum of two even numbers is also even) as an example.

Variance component models

For the i^{th} student from the j^{th} school, the probability that this student will choose a particular option h ($h = 1, .., 6$) is $\pi_{ij}^{(h)}$, and $\sum_{h=1}^{6} \pi_{ij}^{(h)} = 1$.

Through a generalisation of the ordinary logit model, we can define a multinomial logit model for the outcomes. A base category t is needed to form the logit function. We chose the proportion choosing the empirical answer as the base group for question A1.

The simplest model, fitting other 5 categories and taking into account the school clustering, is a two-level one written as,

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(t)}}\right) = \beta_0^{(s)} x_0^{(s)} + u_{0j}^{(s)}, \qquad s = 1, .., 5 \qquad (1)$$

$$x_0^{(s)} = \left\{\begin{array}{lll} x_0^{(1)} = 1 & if & Exhaustive \\ x_0^{(2)} = 1 & if & Formal(c) \\ x_0^{(3)} = 1 & if & Formal(i) \\ x_0^{(4)} = 1 & if & Narrative \\ x_0^{(5)} = 1 & if & Visual \end{array}\right\} 0, Otherwise$$

$\beta_0^{(s)}$ are the parameter estimates associated with the logarithm of the ratios between each of these category and the base category. A positive estimate means a higher likelihood of choosing category s than the base category. In the model there are five parts to represent each of the five outcomes.

Based on (27), we can predict the $\pi_{ij}^{(s)}$ as

$$\pi_{ij}^{(s)} = \frac{\exp(\beta_0^{(s)} x_0^{(s)} + u_{0j}^{(s)})}{1 + \sum_{h=1}^{t-1} \exp(\beta_0^{(s)} x_0^{(s)} + u_{0j}^{(s)})} \quad (2)$$

The term $u_{0j}^{(s)}$ measures the random departure (the residual) of the school mean from the overall $\beta_0^{(s)}$ of the category s and we can obtain the school estimates for each category as

$$u_{0j}^{(s)} \sim N(0, \Sigma) \quad (3)$$

From (1), we have the predicted ratio of category s over t of school j as the exponential term of the overall ratio multiplied by that of the school departure term.

$$\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(t)}} = e^{\beta_0^{(s)}} \times e^{u_{0j}^{(s)}} \quad (4)$$

where the variance-covariances of the departure terms of the five categories are

$$\Sigma = \mathrm{var}(u_{0j}^{(s)}) = \begin{pmatrix} \sigma_{u0}^{2(2)} & & & & \\ \sigma_{u0}^{(2,3)} & \sigma_{u0}^{2(3)} & & & \\ \sigma_{u0}^{(2,4)} & \sigma_{u0}^{(3,4)} & \sigma_{u0}^{2(4)} & & \\ \sigma_{u0}^{(2,5)} & \sigma_{u0}^{(3,5)} & \sigma_{u0}^{(4,5)} & \sigma_{u0}^{2(5)} & \\ \sigma_{u0}^{(2,6)} & \sigma_{u0}^{(3,6)} & \sigma_{u0}^{(4,6)} & \sigma_{u0}^{(5,6)} & \sigma_{u0}^{2(6)} \end{pmatrix} \quad (5)$$

The parameters in (5) are estimated at school level and are known as random effects of parameters $\beta_0^{(s)}$.

To estimate the model, the second order Taylor series approximation and the penalised quasi-likelihood procedure (PQL) are used by means of MLn (Goldstein, 1996). The hypothesis around the main effects can be tested by the Ward χ^2 statistic. The approximative χ^2 test can be also used to test for differences among $\beta_0^{(s)}$. (See Goldstein, 1995 for details).

This simple model produces the following estimates:

Note:
A1-2: Duncan's exhaustive option
A1-5: Arthur's formal correct option
A1-6: Eric's formal incorrect option
A1-7: Ceri's narrative option
A1-8: Yvonne's visual option

Fixed (main effects)

	ESTI (S.E.)	Significant
A1-2	0.228 (0.067)	P<0.01
A1-5	-0.753 (0.136)	P<0.01
A1-6	-2.279 (0.162)	P<0.01
A1-7	-0.287 (0.112)	P<0.01
A1-8	-0.428 (0.081)	P<0.01

Random (Variance-covariance estimates at school level)

	A1-2	A1-5	A1-6	A1-7	A1-8
A1-2	0.20(0.06)				
A1-5	0.36(0.09)	1.22(0.25)			
A1-6	0.21(0.10)	0.88(0.23)	0.62(0.34)		
A1-7	0.37(0.08)	0.84(0.17)	0.71(0.18)	0.83(0.17)	
A1-8	0.11(0.05)	0.32(0.11)	0.28(0.13)	0.19(0.09)	0.26(0.09)

In line with the descriptive statisitcs (see Figure 1a), the main effects suggest that in general more students choose the exhaustive option than the empirical, while fewer students choose either formal presentation, the narrative or the visual than the empirical. The least popular choice by students is the formal incorrect option. The predicted proportions for the each choice can be calculated, based on equation (2) but ignoring the departure terms. For example the predicted proportion choosing the exhaustive is calculated as:

$$\pi^{(2)} = \frac{\exp(0.228)}{1+\exp(0.228)+\exp(-0.753)+\exp(-2.279)+\exp(-0.287)+\exp(-0.428)} = 0.297,$$

and the predicted proportions for the other four are 0.111, 0.024, 0.177 and 0.154 respectively. The base is 0.237.

At school level we have a full variance-covariance structure for the random effects of the $\beta_0^{(s)}$. This suggests that all five ratios are varying from school to school. The distribution for choice on exhaustive form is estimated as $0.228 \pm 2\times\sqrt{0.20} = (-0.666, 1.122)$ for about 95% of the schools. Taking the exponential of the range gives the lower and upper ratios among schools as 0.514 and 3.07 respectively around the overall ratio 1.256. The same calculation can be done for the other four categories.

Modelling the fixed effects of level 1 and level 2 variables

To examine how the proportions choosing different proof forms can be affected by other variables, we need to add these variables into model (1). Suppose we have the variable x_1 representing a Level 1 variable, say sex of student or Key Stage 3 test score, and x_2 representing a Level 2 variable, say whether the school is selective or not, as in (6)

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(t)}}\right) = \beta_0^{(s)} x_0^{(s)} + \beta_1^{(s)} x_{1ij}^{(s)} + \beta_2^{(s)} x_{2j}^{(s)} + u_{0j}^{(s)}, \qquad s=1,\ldots,5 \qquad (6)$$

(32)

We are now estimating a set of parameters associated with x_1 and another set of parameters associated with x_2 from this model. If x_1 is coded as boy = 0 and girl = 1, the predicted overall proportions of the form s for a boy and a girl conditional on x_2 are respectively,

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(t)}}\right)_b = \beta_0^{(s)} x_0^{(s)} \qquad (6.1) \qquad \log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(t)}}\right)_g = \beta_0^{(s)} x_0^{(s)} + \beta_1^{(s)} x_{1ij}^{(s)} \qquad (6.2)$$

Subtracting (6.1) from (6.2), we obtain the equation (6.3), the explanation of the estimated $\beta_1^{(s)}$ as the log odds-ratio of the form s between girls and boys. A positive estimate for a presentation form indicates a preference of girls over boys on this particular form of presentation.

$$\log\left(\frac{\pi_{ij_g}^{(s)} / \pi_{ij_g}^{(t)}}{\pi_{ij_b}^{(s)} / \pi_{ij_b}^{(t)}}\right) = \beta_1^{(s)} \qquad (6.3)$$

For a continuous x_1, its estimate predicts the increase or decrease of the $\log(\pi_{ij}^{(s)} / \pi_{ij}^{(t)})$ as x_1 increases by one unit. Further variables can be added to the model at both Level 1 and Level 2 as required.

REFERENCES

Department of Education and Science (1982). Mathematics counts: Report of the Committee of Enquiry into the Teaching of Mathematics in School. London: Her Majesty's Stationery Office.

Department for Education (1994). Mathematics in the National Curriculum. London: HMSO.

Goldstein, H. (1995). Multilevel Statistical Models. Edward Arnold.