Appendix Z

Longitudinal Proof Project

Multilevel models of the first year's data

Geoffrey Woodhouse

Institute of Education
University of London

August, 2001

CONTENTS

Proof Project: Multilevel models of the first year's data

1 Introduction

In this report, we model:

1. Total scores for constructive proof in Geometry and Algebra (combined with Logic)
2. Validity rating (VR) scores in Geometry (G3) and Algebra (A3)
3. Scores for individual constructive proofs in Algebra, Logic, and Geometry
4. Probabilities of different choices for own approach in G3 and A3
5. Probabilities of different choices for best mark in G3 and A3

For 1, 2, and 4, we develop models both to compare schools and to explain what is going on at student level. For 3 and 5 we develop explanatory models only.

2 Total scores for Geometry and Algebra constructive proof

In this section, we model the total raw scores for Algebra (A1, A2, A4, L1) and Geometry (G1, G2, G4) as a bivariate response. In a later section we normalise these total scores across the whole sample, in order to compare effect sizes across several outcomes. The main purpose of this kind of modelling is to obtain estimates of the correlations between school effects on the two outcomes, if any, and also to study correlations at student level. A subsidiary purpose is to obtain more precise estimates of all effects, both fixed and random.

Data were available on 2799 students in 115 classes spread over 63 schools. Not all classes had teacher data, and not all the student data were complete.

2.1 Models for comparing schools

The following data were considered for inclusion:

school-level:
administration (County, Voluntary (VA/VC), or other)
whether 11-16 or 11-18
% A*-C at GCSE
gender
area (one of three)
year-8 size (<180, 180-219, >=220)
% to be entered at lev6-8 at KS3 (<20, 20-49, 50-89, >=90)
GCSE syllabus
maths textbook or scheme in use
minutes of maths per week(<=175, 180, >180)
existence of maths club

class-level:
teacher gender
teacher years of experience
teacher age
teacher degree (including whether or not a maths degree)
teacher PGCE or Cert
teacher HE
teacher involvement in INSET, whether school, LEA, college, distance, or other
teacher membership of a professional assoc.
teacher knowledge/use of software

teacher extra-curricular activity

student-level:
gender
age in months
score on baseline test
scores on each section of the proof assessment (the *response*)
(note that other student scores on the proof assessment were associated with differences in constructive proof scores, but these are not relevant for comparing schools)

Of the above, statistically significant main effects were found for:

school-level:
% A*-C at GCSE
gender
maths textbook or scheme in use
minutes of maths per week

class-level:
teacher gender

student-level:
gender
score on baseline test

None of the other variables was found to have a statistically significant effect on either of the outcome scores. In particular, whether a school was 11-18 or 11-16 had no effect. The school-gender 'effect' was apparently negative on algebra scores in girls-only schools. There is little *prima facie* reason to expect girls to perform less well in algebraic proof when they are in girls' schools, when there is no such effect on their Geometry scores. The teacher-gender 'effect' suggested that girls perform better at geometrical proof than boys, but only when taught by a woman teacher. This also has little *prima facie* justification. With so many variables discretion is needed, and reliance solely on statistical significance as a criterion is not appropriate. Thus, we have ignored these two effects. (The children in the girls' schools in this sample have tended to do less well in algebraic proof than the girls in the other schools.)

The apparent effect of minutes of maths per week was negative on both scores. This is counter-intuitive. Something else is likely to be going on in the schools that happen to report offering more than 180 minutes per week of maths, and accordingly we exclude this effect.

We are left with two effects at school level, two at student level, and none at class or teacher level.

In our first model, Model 0, the only fixed effect included, apart from the intercept ('geo_cons', 'alg_cons'), is gender ('geo_girl', 'alg_girl'). We allow three levels of variation, with students at level 2, classes at level 3, and schools at level 4. Level 1 is used to distinguish between the responses in Algebra and Geometry. This is a standard way to set up a multivariate multilevel model (see, for example, Goldstein, 1995, Chapter 4). The reason for including this very basic model is to check whether there is a gender effect, and whether there is detectable variation in the outcomes at class level, in a model unadjusted for baseline score.

In the tabulation of the fixed estimates for this and subsequent models, the prefix 'geo_' indicates an effect on the score in Geometry; 'alg_' indicates an effect on the score in Algebra.

Model 0

The fixed-part estimates from this model may be expressed by means of the two equations:

$$\text{predicted geometry score} = 8.069 - 0.065 girl,$$
$$\text{predicted algebra score} = 9.661 + 0.031 girl,$$

thus, the predicted score for the base group (boys) in Geometry is 8.069 (s.e. 0.226) and in Algebra it is 9.661 (s.e. 0.264). The gender coefficients are statistically non-significant, as is shown in the full tabulation below:

PARAMETER	ESTIMATE	S. ERROR
geo_girl	-0.06458	0.1222
alg_girl	0.03115	0.155
geo_cons	8.069	0.2263
alg_cons	9.661	0.2644

Thus, there is *no* statistically significant gender effect on either outcome score when the model makes no adjustment for baseline score.

The residual variance/covariance matrices for the outcome scores at the three levels, school, class, and student, have the following estimates (covariances have been converted to correlations for convenience):

School level

	Geo	*Alg*
Geo	2.47	
Alg	$r = .88$	3.07

Class level

	Geo	*Alg*
Geo	.41	
Alg	$r = .88$	.84

Student level

	Geo	*Alg*
Geo	9.73	
Alg	$r = .34$	15.62

The full tabulation is:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ^2_w(geo_cons)	2.467	0.54	1
σ_w(geo_cons, alg_cons)	2.42	0.5773	0.88
σ^2_w(alg_cons)	3.07	0.7373	1
Class			
σ^2_v(geo_cons)	0.4055	0.1729	1
σ_v(geo_cons, alg_cons)	0.5147	0.1965	0.881
σ^2_v(alg_cons)	0.8417	0.3178	1
Student			
σ^2_u(geo_cons)	9.726	0.2652	1
σ_u(geo_cons, alg_cons)	4.218	0.2516	0.342
σ^2_u(alg_cons)	15.62	0.4267	1

In the above table, as in all other tables describing random effects, subscripts *u*, *v*, and *w* are used to distinguish levels of variance and covariance. The variable(s) to which a (co)variance refers are/is given in parentheses. The suffix 'cons' indicates an intercept term. There is statistically significant, though small, residual variation at class level, and high residual correlation ($r \cup .88$) at both school and class level between the two outcomes. Thus, in this very simple model, schools that perform above the average in algebra are predicted to do so in geometry also, and vice versa. Furthermore, class effects on the two subjects, within schools, are predicted to be similar. A student, however,

who performs above the expectation for her class in algebra has only a slight tendency to perform above expectation in geometry also. In percentages, the residual variances at school, class, and student levels, are in the ratio 20:3:77 for Geometry and 16:4:80 for Algebra.

We have included the baseline score as a predictor in all subsequent models. This means that we are assessing the effects of the other variables, together with any random effects at student, class, and school level, on the proof scores of children with apparently equal attainment in other areas of maths. Thus, for the purpose of school comparison, we are treating the baseline score as an intake measure.

Once baseline score is included, no statistically significant variation remains at class level within school. Thus, we have two levels of variation, students at level 2 within schools at level 3, with level 1 used as before to distinguish between the responses in Algebra and Geometry.

We find that, in general, there is a significant student-gender effect, in favour of girls, which we include in all models for school comparison. This is in addition to the baseline score effect. There is no detectable interaction between gender and baseline score, in other words, the advantage to girls appears to be equal for all values of the baseline score. The base category for gender in the fixed part of all models is boys. In Models 1 and 2, we allow the student-gender effect to vary at both school and student level. This is equivalent to allowing schools to be differentially 'effective' for girls and boys, and also allows that individually boys may vary differently from girls about their predicted scores. In Models 3 and 4 the variance at school level is pooled between girls and boys, in other words, school 'effects' are assumed to be the same for girls as for boys. No school- or student-level variation has been found in the baseline score effect in any of our models.

There is a question whether, in a model for comparing schools, the school % A*-C should be included. Models 1 and 3 exclude this predictor; Models 2 and 4 include it. Undoubtedly, however, the textbook chosen by the school should *not* be included in a model for comparing schools, though it does appear when we extend the model to explain what is going on at student level (Model 5).

Model 1

In this model, the fixed-part predictors are gender and baseline score, standardised to have a mean of zero and a standard deviation of 1. The fixed part of the model for geometry scores is estimated to be:

$$\text{predicted geometry score} = 7.58 + 1.52base + 0.2305base^2 + 0.06163base^3 + 0.4148girl\ ,$$

a cubic in the baseline score, with an additional predicted benefit of 0.4148 raw-score point for the girls over the boys.

The corresponding model for algebra scores is

$$\text{predicted algebra score} = 8.944 + 1.846base + 0.4692base^2 + 0.1571base^3 + 0.5519girl,$$

also a cubic in the baseline score with, this time, an additional predicted benefit of 0.5519 raw-score point for the girls over the boys.

The full tabulation of the fixed part, showing the standard errors, is:

PARAMETER	ESTIMATE	S. ERROR
geo_girl	0.4148	0.125
alg_girl	0.5519	0.1638
geo_cons	7.58	0.1537
alg_cons	8.944	0.205
geo_base	1.52	0.09645
alg_base	1.846	0.1209
geo_base2	0.2305	0.06097
alg_base2	0.4692	0.07647
geo_base3	0.06163	0.0314
alg_base3	0.1571	0.0394

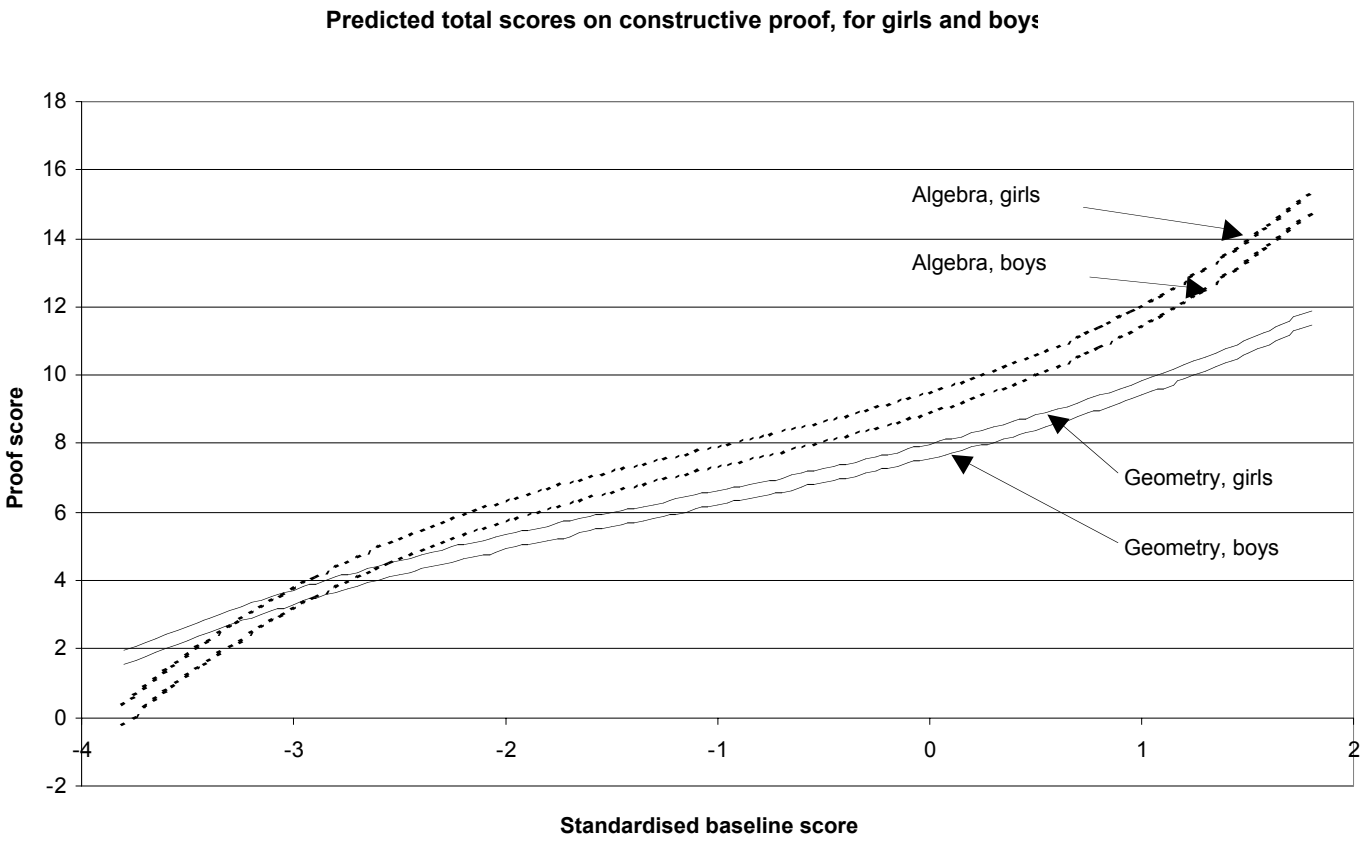
The relationship described by each equation is illustrated in the first of the two graphs on the following pages. The second graph shows the relationship of the outcome scores with the raw baseline score. The mean of the raw baseline scores is 15.30 and the SD is 3.79. The distribution of the baseline scores is illustrated below.

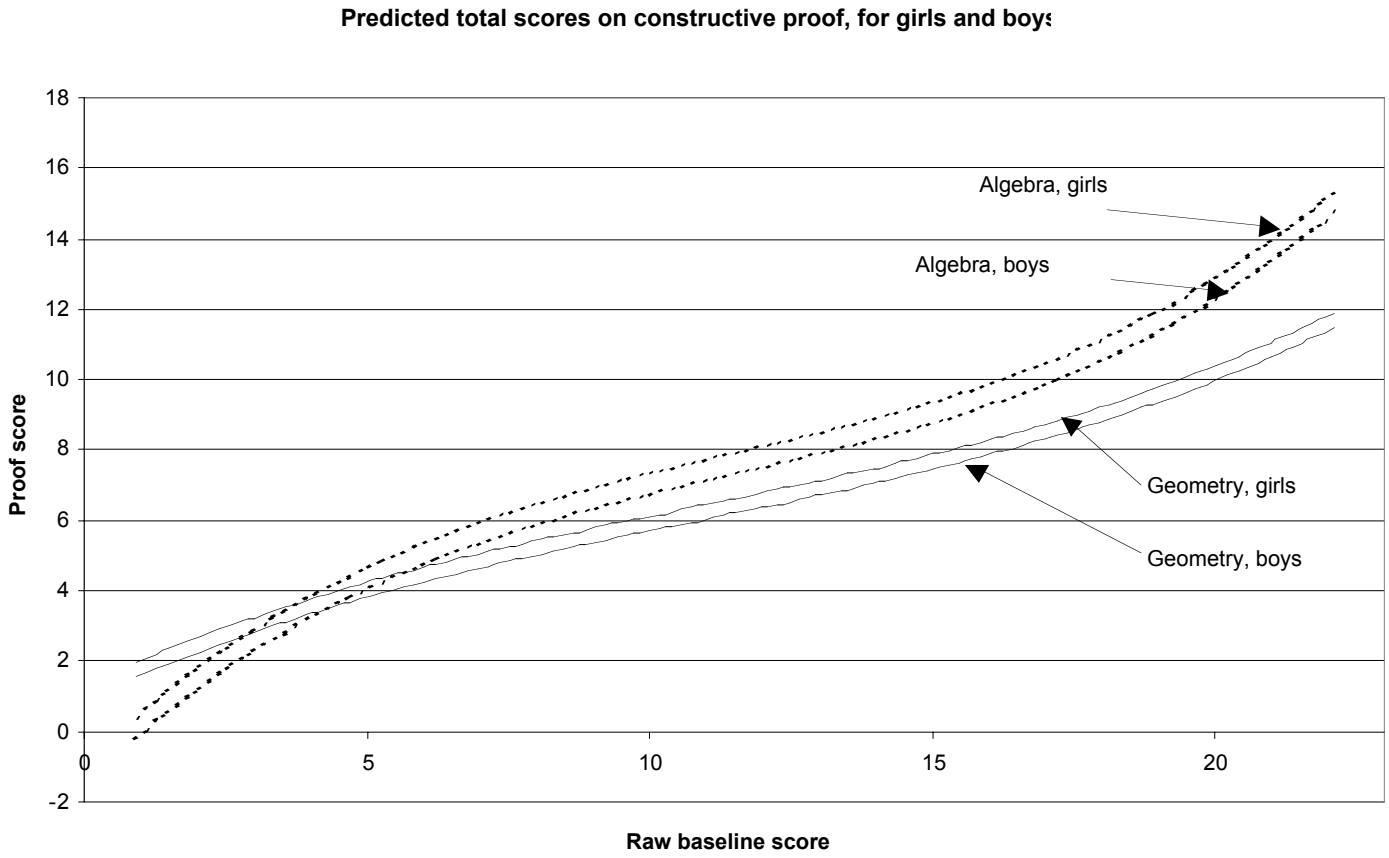
Distribution of baseline scores

Each * = up to 6 cases

Lower limit	N	
1.000	1	*
2.000	2	*
3.000	2	*
4.000	11	**
5.000	15	***
6.000	25	*****
7.000	30	*****
8.000	48	********
9.000	64	**********
10.00	93	**********
11.00	147	**********
12.00	165	**********
13.00	221	**********
14.00	210	**********
15.00	236	**********
16.00	289	**********
17.00	260	**********
18.00	269	**********
19.00	226	**********
20.00	160	**********
21.00	131	**********
22.00	58	**********

There were 136 students whose baseline score was missing. Hence the model is based on the remaining 2663 students.


Predicted total scores on constructive proof, for girls and boys
Proof score
Standardised baseline score
Algebra, girls
Algebra, boys
Geometry, girls
Geometry, boys
18
16
14
12
10
8
6
4
2
0
-2
-4
-3
-2
-1
0
1
2


Predicted total scores on constructive proof, for girls and boys
Proof score
Raw baseline score
Algebra, girls
Algebra, boys
Geometry, girls
Geometry, boys
18
16
14
12
10
8
6
4
2
0
-2
0
5
10
15
20

The random part of Model 1 can be expressed as follows:

School level (variances on the diagonal; correlations elsewhere)

		Girls		*Boys*	
		Geo	*A*lg	*Geo*	*A*lg
Girls	*Geo*	.902			
	*A*lg	$r = .53$	.768		
Boys	*Geo*	$r = .92$	$r = .58$	.853	
	*A*lg	$r = .64$	$r = .92$	$r = .54$	1.62

Student level (variances on the diagonal; correlations elsewhere)

		Girls		*Boys*	
		Geo	*A*lg	*Geo*	*A*lg
Girls	*Geo*	7.91			
	*A*lg	$r = .19$	12.51		
Boys	*Geo*	0	0	8.42	
	*A*lg	0	0	$r = .25$	13.28

The full tabulation is:

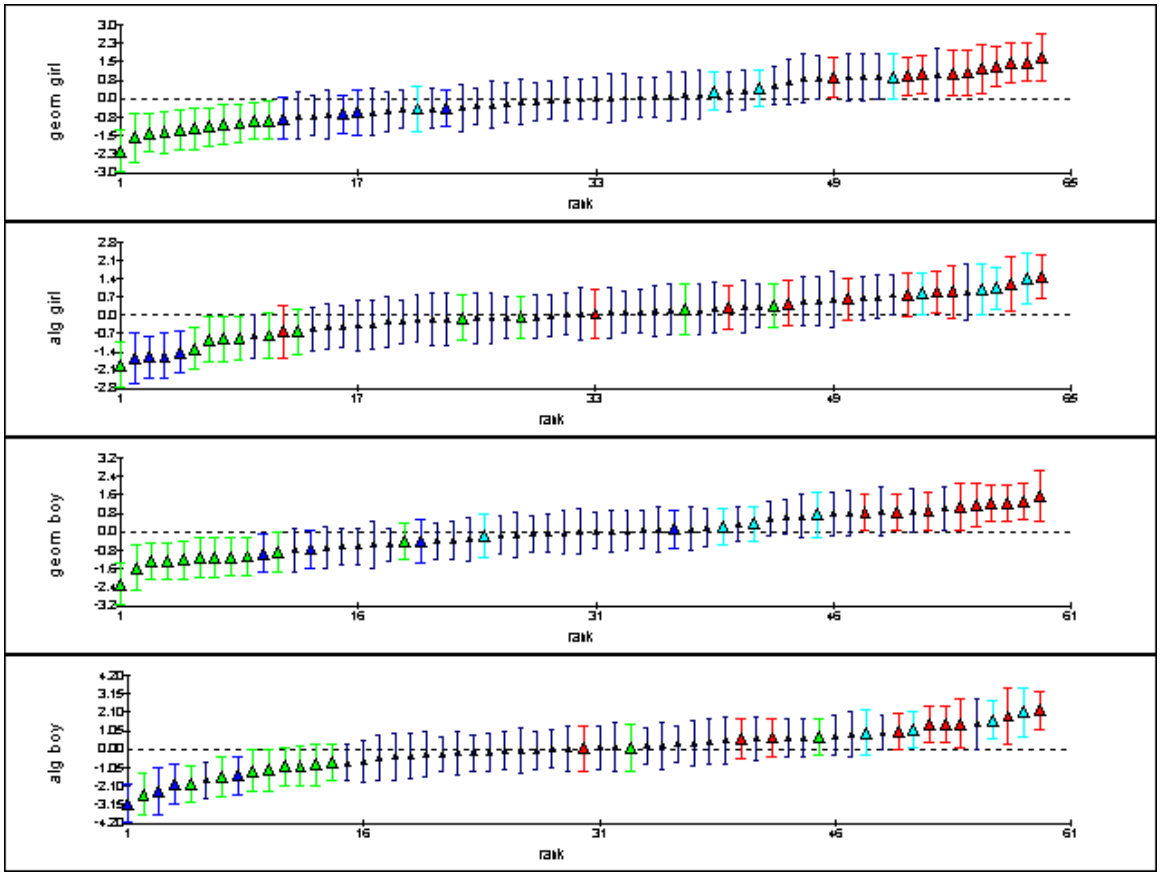
PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(geo_girl)	0.9015	0.2283	1
σ_v(alg_girl,geo_girl)	0.441	0.1795	0.53
σ_v^2(alg_girl)	0.768	0.2431	1
σ_v(geo_boy,geo_girl)	0.8095	0.1921	0.923
σ_v(geo_boy,alg_girl)	0.4684	0.1782	0.579
σ_v^2(geo_boy)	0.8531	0.2286	1
σ_v(alg_boy,geo_girl)	0.7779	0.2399	0.643
σ_v(alg_boy,alg_girl)	1.031	0.2609	0.924
σ_v(alg_boy,geo_boy)	0.6295	0.2376	0.535
σ_v^2(alg_boy)	1.621	0.4136	1
Student			
σ_u^2(geo_girl)	7.908	0.3105	1
σ_u(alg_girl,geo_girl)	1.861	0.2812	0.187
σ_u^2(alg_girl)	12.51	0.4915	1
σ_u^2(geo_boy)	8.415	0.337	1
σ_u(alg_boy,geo_boy)	2.61	0.3092	0.247
σ_u^2(alg_boy)	13.28	0.5335	1

The very high correlations between girls' and boys' scores (within subject) at school level mean that, for example, a school that has a high residual for boys is very likely to have a high residual

also for girls in the same subject. There is moderate, but statistically significant, correlation between residuals for geometry and algebra between schools.

It appears from the tables that schools are more variable in their boys' residual performance at algebra (compared to predictions) than in their girls' performance: but this difference does not reach statistical significance. Elsewhere, the variances of boys and girls within subject are similar. A correlation between boys and girls is not meaningful at student level (i.e. within a student), and correlations within students between their residual performance in algebra and in geometry (compared to predictions) are weak. A student who performs better than expected for their school in, say, algebra (after adjustment for their gender and baseline score and the school's residual) is not especially likely to perform better than expected for their school in geometry also.

The following chart plots school-level residuals against their ranks, with error bars corresponding to 1.96 SD. Thus, an error bar wholly above the dotted line corresponds to a school that is performing above the mean predicted by the model, with 95% confidence.



There are 63 schools in the sample, and all have girl students. Four of the schools are girls-only, hence only 59 schools feature in the chart for boys' residuals. Schools performing above the mean in geometry for girls have been highlighted in red. It is clear that these tend also to perform above the mean in geometry for boys but not necessarily in algebra for either girls or boys. Low performing schools in geometry for girls are highlighted in green, and these tend also to perform below the mean in geometry for boys. The remaining high-performing schools in algebra for girls are highlighted in cyan: these include the remaining high-performing schools in algebra for boys, except school 26. The remaining low-performing schools in algebra for girls are highlighted in blue and these include the remaining low-performing schools in algebra for boys, except school 6. One school, while ranking in the top ten for all four residuals, is not highlighted in the chart since its

residuals cannot be distinguished from the mean with 95% confidence. This is school 53. It has only 24 students in the sample, making it less likely that statistically significant residuals will be detected for it.

Tables of the school residual ranks now follow. Schools are ranked from 63 downwards for girls and from 59 downwards for boys. (A high-numbered rank indicates good performance.)

School residual ranks from Model 1:

School	geo_girl	alg_girl	geo_boy	alg_boy
1	48	43	N/A	N/A
2	43	35	41	35
3	6	13	5	12
4	9	45	4	46
5	40	49	N/A	N/A
6	36	10	43	6
7	60	63	57	59
8	22	29	21	25
9	19	51	22	44
10	39	41	36	39
11	4	24	11	13
12	54	50	51	51
13	34	19	38	17
14	23	5	13	8
15	2	39	2	34
16	17	2	20	3
17	52	40	47	40
18	28	18	23	21
20	55	42	49	43
21	44	60	40	57
22	35	37	30	38
23	16	3	37	1
24	61	56	59	54
25	42	44	35	47
26	51	48	54	42
27	32	14	32	15
28	29	30	28	26
29	12	4	10	4
30	63	61	N/A	N/A
31	57	12	N/A	N/A
32	24	23	24	19
33	46	47	45	48

School	geo_girl	alg_girl	geo_boy	alg_boy
34	18	26	14	30
35	3	6	3	5
36	1	1	1	2
37	26	22	27	20
38	33	21	33	23
39	11	9	9	11
40	15	38	17	33
41	13	25	12	27
42	7	11	7	9
43	14	27	15	24
44	25	16	18	22
45	20	31	16	37
46	5	7	8	7
47	27	36	29	29
48	21	59	25	49
49	8	8	6	10
50	37	20	26	28
51	50	32	48	36
52	30	52	39	45
53	56	58	50	56
54	31	15	34	16
55	47	34	52	32
56	59	57	55	55
57	49	46	53	41
58	10	28	19	14
59	62	54	58	53
60	41	55	42	52
61	58	33	56	31
62	45	53	44	50
63	38	17	31	18
64	53	62	46	58

Schools ranked according to their residuals for girls' geometry and girls' algebra (Model 1):

Ranked according to girls' geometry				
School	geo_girl	alg_girl	geo_boy	alg_boy
30	63	61	N/A	N/A
59	62	54	58	53
24	61	56	59	54
7	60	63	57	59
56	59	57	55	55
61	58	33	56	31
31	57	12	N/A	N/A
53	56	58	50	56
20	55	42	49	43
12	54	50	51	51
64	53	62	46	58
17	52	40	47	40
26	51	48	54	42
51	50	32	48	36
57	49	46	53	41
1	48	43	N/A	N/A
55	47	34	52	32
33	46	47	45	48
62	45	53	44	50
21	44	60	40	57
2	43	35	41	35
25	42	44	35	47
60	41	55	42	52
5	40	49	N/A	N/A
10	39	41	36	39
63	38	17	31	18
50	37	20	26	28
6	36	10	43	6
22	35	37	30	38
13	34	19	38	17
38	33	21	33	23
27	32	14	32	15
54	31	15	34	16
52	30	52	39	45
28	29	30	28	26
18	28	18	23	21
47	27	36	29	29
37	26	22	27	20
44	25	16	18	22
32	24	23	24	19
14	23	5	13	8
8	22	29	21	25
48	21	59	25	49
45	20	31	16	37
9	19	51	22	44
34	18	26	14	30
16	17	2	20	3
23	16	3	37	1
40	15	38	17	33

Ranked according to girls' algebra				
School	geo_girl	alg_girl	geo_boy	alg_boy
7	60	63	57	59
64	53	62	46	58
30	63	61	N/A	N/A
21	44	60	40	57
48	21	59	25	49
53	56	58	50	56
56	59	57	55	55
24	61	56	59	54
60	41	55	42	52
59	62	54	58	53
62	45	53	44	50
52	30	52	39	45
9	19	51	22	44
12	54	50	51	51
5	40	49	N/A	N/A
26	51	48	54	42
33	46	47	45	48
57	49	46	53	41
4	9	45	4	46
25	42	44	35	47
1	48	43	N/A	N/A
20	55	42	49	43
10	39	41	36	39
17	52	40	47	40
15	2	39	2	34
40	15	38	17	33
22	35	37	30	38
47	27	36	29	29
2	43	35	41	35
55	47	34	52	32
61	58	33	56	31
51	50	32	48	36
45	20	31	16	37
28	29	30	28	26
8	22	29	21	25
58	10	28	19	14
43	14	27	15	24
34	18	26	14	30
41	13	25	12	27
11	4	24	11	13
32	24	23	24	19
37	26	22	27	20
38	33	21	33	23
50	37	20	26	28
13	34	19	38	17
18	28	18	23	21
63	38	17	31	18
44	25	16	18	22
54	31	15	34	16

43	14	27	15	24
41	13	25	12	27
29	12	4	10	4
39	11	9	9	11
58	10	28	19	14
4	9	45	4	46
49	8	8	6	10
42	7	11	7	9
3	6	13	5	12
46	5	7	8	7
11	4	24	11	13
35	3	6	3	5
15	2	39	2	34
36	1	1	1	2

27	32	14	32	15
3	6	13	5	12
31	57	12	N/A	N/A
42	7	11	7	9
6	36	10	43	6
39	11	9	9	11
49	8	8	6	10
46	5	7	8	7
35	3	6	3	5
14	23	5	13	8
29	12	4	10	4
23	16	3	37	1
16	17	2	20	3
36	1	1	1	2

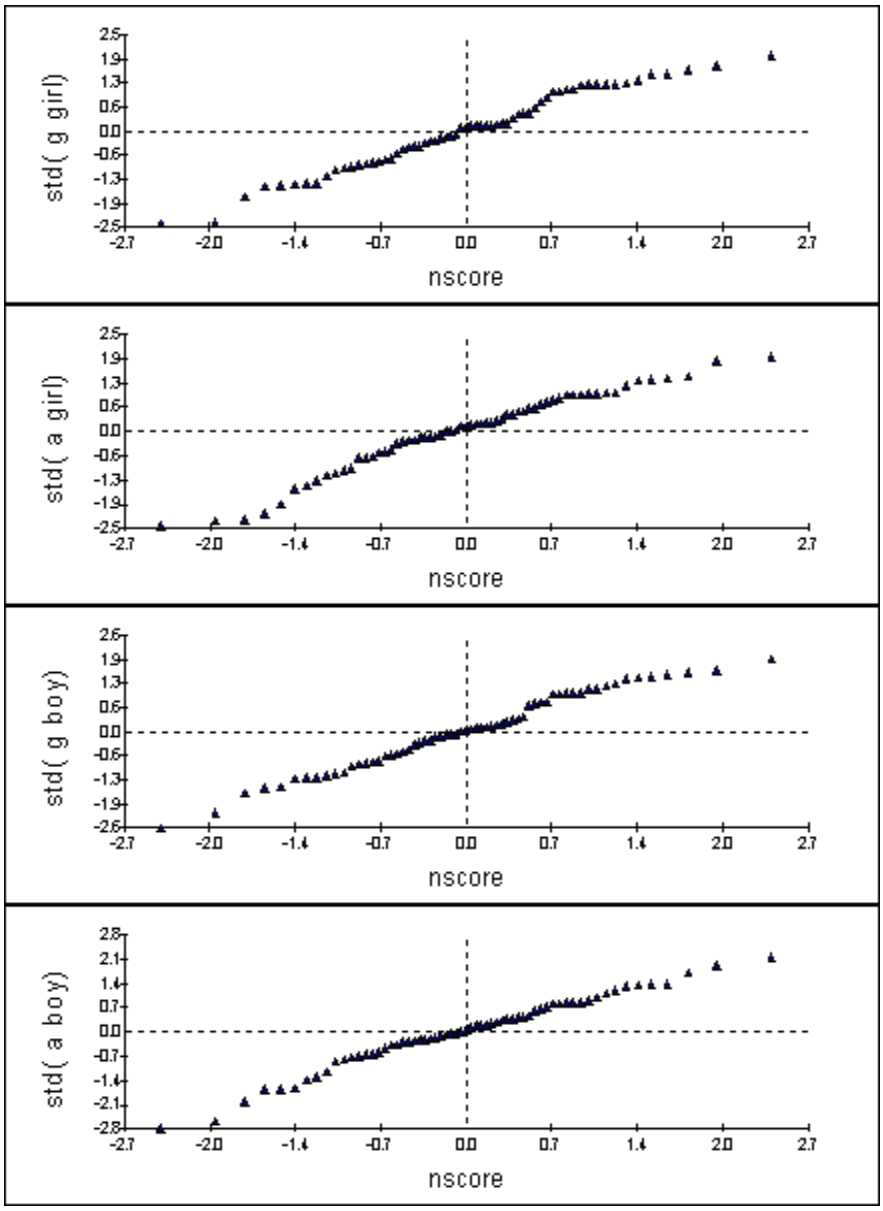
Above the mean for girls' Geometry: 30, 59, 24, 7, 56, 61, 31, 20, 12, 57
Above the mean for boys' Geometry: 24, 59, 7, 61, 56, 57, 12, 20
Below the mean for girls' Geometry: 36, 15, 35, 11, 46, 3, 42, 49, 4, 58
Below the mean for boys' Geometry: 36, 15, 35, 4, 3, 49, 46, 39, 29

Above the mean for girls' Algebra: 7, 64, 30, 21, 48, 60
Above the mean for boys' Algebra: 7, 64, 21, 56, 24, 26, 59, 60
Below the mean for girls' Algebra: 36, 16, 23, 29, 14, 35, 46, 39
Below the mean for boys' Algebra: 23, 36, 16, 29, 35, 6, 46, 14

Large differences between rankings on girls' Geometry and Algebra: 31, 48, 15, 4, 9
Large differences between rankings on boys' Geometry and Algebra: 4, 6, 23, 15

We also plotted standardised diagnostic school-level residuals against their normal scores (see below). These are not ideal, and reflect problems in the scoring of the outcome.

Standardised diagnostic school-level residuals against their normal scores, Model 1



Model 2

An alternative model for comparing schools includes as an additional fixed effect the school % GCSE pass rate at A*-C, a historical measure which might be regarded as a compositional effect on the sample students' performance in proof. The other effects in this model, including school-level residuals, are conditional on this effect. Thus, if this model is used to compare schools, the schools are being compared for the additional effects they are having on students' performance in proof, over and above what may underlie historical performance at GCSE.

The distribution of the %A*-C variable across students is illustrated below:

```
Each * = up to 8 cases
    Lower
    limit       N

        15.00     29 : ****
        18.00     48 : ******
        21.00      0 :
        24.00    132 : ******************
        27.00     28 : ****
        30.00    119 : ***************
        33.00     62 : ********
        36.00     58 : ********
        39.00    221 : ****************************
        42.00    113 : ***************
        45.00    296 : ************************************
        48.00    154 : ********************
        51.00    388 : ************************************************
        54.00    205 : ***************************
        57.00    150 : ********************
        60.00    284 : ************************************
        63.00     44 : ******
        66.00    221 : ****************************
        69.00     64 : ********
        72.00     95 : ************
        75.00      0 :
        78.00      0 :
        81.00     31 : ****
        84.00     32 : ****
        87.00      0 :
        90.00      0 :
        93.00      0 :
        96.00      0 :
        99.00     25 : ****
        102.0      0 :
```

The raw mean is 50.7 and the SD is 14.7.

The fixed part of Model 2 is:

$$\begin{aligned}\text{predicted geometry score} &= 7.565 + 1.485\,base + 0.2248\,base^2 + 0.06342\,base^3 \\ &\quad + 0.397\,girl + 0.5249 \leftarrow\!\!\!\!\!\!\% A\_C, \\ \text{predicted algebra score} &= 8.919 + 1.814\,base + 0.4633\,base^2 + 0.1586\,base^3 \\ &\quad + 0.5513\,girl + 0.4627 \leftarrow\!\!\!\!\!\!\% A\_C,\end{aligned}$$

where %A_C has been standardised to have a mean of zero and a SD of 1 over all students. This variable has a positive effect, with a predicted difference of about half a raw-score point in both Geometry and Algebra, per standard deviation of the variable.

The fixed-part parameters are tabulated, with their standard errors, on the next page.

PARAMETER	ESTIMATE	S. ERROR
geo_girl	0.397	0.1254
alg_girl	0.5513	0.1643
geo_cons	7.565	0.1421
alg_cons	8.919	0.1939
geo_base	1.485	0.09678
alg_base	1.814	0.1218
geo_base2	0.2248	0.06078
alg_base2	0.4633	0.07634
geo_base3	0.06342	0.0313
alg_base3	0.1586	0.03933
geo_%_a_c	0.5249	0.103
alg_%_a_c	0.4627	0.1221

Turning to the random part of Model 2, we find that including the additional predictor in the fixed part reduces the school-level variation. Schools' residual performance in algebra is significantly more variable for boys than for girls once the effect of schools' % A*-C is included. Two of the correlations at school level between performance in Geometry and in Algebra fail to reach statistical significance, but we retain them as the others are significant and all are indicative. There is little change from Model 1 in the random part at student level.

School level (variances on the diagonal; correlations elsewhere)

		Girls		*Boys*	
		Geo	*A*lg	*Geo*	*A*lg
Girls	*Geo*	.553			
	*A*lg	$r = .29\,ns$	.536		
Boys	*Geo*	$r = .88$	$r = .38\,ns$	.638	
	*A*lg	$r = .53$	$r = .91$	0.41	1.34

The full tabulation of the random part of Model 2 is:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(geo_girl)	0.5528	0.1657	1
σ_v(alg_girl,geo_girl)	0.1583	0.1326	0.291
σ_v^2(alg_girl)	0.5362	0.2008	1
σ_v(geo_boy,geo_girl)	0.5248	0.1432	0.884
σ_v(geo_boy,alg_girl)	0.2212	0.1425	0.378
σ_v(geo_boy,geo_boy)	0.6376	0.1899	1
σ_v(alg_boy,geo_girl)	0.4572	0.1846	0.531
σ_v(alg_boy,alg_girl)	0.7708	0.2162	0.909
σ_v(alg_boy,geo_boy)	0.3749	0.1964	0.405
σ_v^2(alg_boy)	1.342	0.3632	1

Student			
σ_u^2(geo_girl)	7.897	0.31	1
σ_u(alg_girl,geo_girl)	1.852	0.2808	0.186
σ_u^2(alg_girl)	12.51	0.4913	1
σ_u^2(geo_boy)	8.412	0.3369	1
σ_u(alg_boy,geo_boy)	2.606	0.3091	0.247
σ_u^2(alg_boy)	13.27	0.5334	1

School residual ranks under Model 2:

School	geo_girl	alg_girl	geo_boy	alg_boy
1	30	18	N/A	N/A
2	43	33	38	30
3	2	14	2	9
4	9	50	5	48
5	46	60	N/A	N/A
6	31	7	45	3
7	60	62	55	58
8	16	26	15	22
9	18	57	21	46
10	28	35	26	31
11	11	43	22	24
12	61	56	57	55
13	21	16	25	13
14	41	5	27	12
15	6	59	7	47
16	29	2	32	2
17	58	36	51	42
18	33	24	28	27
20	53	29	46	36
21	27	51	16	51
22	37	41	33	41
23	15	1	35	1
24	62	55	58	56
25	35	39	29	43
26	42	31	43	29
27	24	9	24	10
28	19	25	17	20
29	34	6	30	7
30	44	37	N/A	N/A
31	57	8	N/A	N/A
32	39	42	41	33
33	40	38	36	35

School	geo_girl	alg_girl	geo_boy	alg_boy
34	13	28	10	26
35	3	4	6	4
36	5	3	3	5
37	51	46	54	44
38	50	40	47	39
39	14	13	8	16
40	32	54	34	49
41	1	15	1	11
42	12	17	11	14
43	4	21	4	17
44	23	20	14	23
45	17	34	12	32
46	7	11	9	6
47	36	44	40	37
48	8	52	13	38
49	26	23	20	21
50	45	30	37	34
51	47	22	42	25
52	22	48	31	40
53	48	47	44	50
54	20	10	18	8
55	54	27	52	28
56	56	49	50	52
57	59	45	56	45
58	10	32	23	18
59	63	53	59	57
60	38	58	39	53
61	49	19	49	19
62	52	61	53	54
63	25	12	19	15
64	55	63	48	59

Schools ranked according to their residuals for girls' geometry and girls' algebra (Model 2):

Ranked according to girls' geometry				
School	geo_girl	alg_girl	geo_boy	alg_boy
59	63	53	59	57
24	62	55	58	56
12	61	56	57	55
7	60	62	55	58
57	59	45	56	45
17	58	36	51	42
31	57	8	N/A	N/A
56	56	49	50	52
64	55	63	48	59
55	54	27	52	28
20	53	29	46	36
62	52	61	53	54
37	51	46	54	44
38	50	40	47	39
61	49	19	49	19
53	48	47	44	50
51	47	22	42	25
5	46	60	N/A	N/A
50	45	30	37	34
30	44	37	N/A	N/A
2	43	33	38	30
26	42	31	43	29
14	41	5	27	12
33	40	38	36	35
32	39	42	41	33
60	38	58	39	53
22	37	41	33	41
47	36	44	40	37
25	35	39	29	43
29	34	6	30	7
18	33	24	28	27
40	32	54	34	49
6	31	7	45	3
1	30	18	N/A	N/A
16	29	2	32	2
10	28	35	26	31
21	27	51	16	51
49	26	23	20	21
63	25	12	19	15
27	24	9	24	10
44	23	20	14	23
52	22	48	31	40
13	21	16	25	13
54	20	10	18	8
28	19	25	17	20
9	18	57	21	46
45	17	34	12	32
8	16	26	15	22
23	15	1	35	1

Ranked according to girls' algebra				
School	geo_girl	alg_girl	geo_boy	alg_boy
64	55	63	48	59
7	60	62	55	58
62	52	61	53	54
5	46	60	N/A	N/A
15	6	59	7	47
60	38	58	39	53
9	18	57	21	46
12	61	56	57	55
24	62	55	58	56
40	32	54	34	49
59	63	53	59	57
48	8	52	13	38
21	27	51	16	51
4	9	50	5	48
56	56	49	50	52
52	22	48	31	40
53	48	47	44	50
37	51	46	54	44
57	59	45	56	45
47	36	44	40	37
11	11	43	22	24
32	39	42	41	33
22	37	41	33	41
38	50	40	47	39
25	35	39	29	43
33	40	38	36	35
30	44	37	N/A	N/A
17	58	36	51	42
10	28	35	26	31
45	17	34	12	32
2	43	33	38	30
58	10	32	23	18
26	42	31	43	29
50	45	30	37	34
20	53	29	46	36
34	13	28	10	26
55	54	27	52	28
8	16	26	15	22
28	19	25	17	20
18	33	24	28	27
49	26	23	20	21
51	47	22	42	25
43	4	21	4	17
44	23	20	14	23
61	49	19	49	19
1	30	18	N/A	N/A
42	12	17	11	14
13	21	16	25	13
41	1	15	1	11

39	14	13	8	16
34	13	28	10	26
42	12	17	11	14
11	11	43	22	24
58	10	32	23	18
4	9	50	5	48
48	8	52	13	38
46	7	11	9	6
15	6	59	7	47
36	5	3	3	5
43	4	21	4	17
35	3	4	6	4
3	2	14	2	9
41	1	15	1	11

3	2	14	2	9
39	14	13	8	16
63	25	12	19	15
46	7	11	9	6
54	20	10	18	8
27	24	9	24	10
31	57	8	N/A	N/A
6	31	7	45	3
29	34	6	30	7
14	41	5	27	12
35	3	4	6	4
36	5	3	3	5
16	29	2	32	2
23	15	1	35	1

Above the mean for girls' Geometry: 59, 24, 12, 7, 57
Above the mean for boys' Geometry: 59, 24, 12, 57, 7, 62
Below the mean for girls' Geometry: 41, 3, 35, 43, 36, 46
Below the mean for boys' Geometry: 41, 3, 36, 43, 4, 35

Above the mean for girls' Algebra: 64, 7, 62
Above the mean for boys' Algebra: 64, 7, 59, 24, 12, 62, 60
Below the mean for girls' Algebra: 23, 16, 36, 35, 14, 29, 6
Below the mean for boys' Algebra: 23, 16, 35, 36, 46

Large differences between rankings on girls' Geometry and Algebra: 15, 31, 48, 4, 9, 14
Large differences between rankings on boys' Geometry and Algebra: 4, 6, 15, 21, 23, 61, 16

Largest changes in rankings from model 1 to model 2:

geo_girl		alg_girl		geo_boy		alg_boy	
sch	change	sch	change	sch	change	sch	change
30	down 19	1	down 25	21	down 24	41	down 14
1	down 18	30	down 24	54	down 16	33	down 13
21	down 17	26	down 17	13	down 13	48	down 11
37	up 25	37	up 24	37	up 25	37	up 22
29	up 22	15	up 20	29	up 19	40	up 15
49	up 18	38	up 19	32	up 17	38	up 14

The next two models are included to illustrate what happens when the random effect at school level is removed from gender and attached instead to the intercept term, which is common to both boys and girls. This is equivalent to pooling schools' performances for their boys and their girls, in other words assuming their 'effects' are the same for either gender. The fixed part changes very little, but a slightly different ranking of schools arises.

Model 3 (as Model 1, but with no random effect of gender at school level)

The model for the fixed part is:

$$\text{predicted geometry score} = 7.583 + 1.526base + 0.2309base^2 + 0.06106base^3 + 0.411girl,$$
$$\text{predicted algebra score} = 8.91 + 1.843base + 0.4747base^2 + 0.1585base^3 + 0.5892girl,$$

that is, very similar to Model 1. Standard errors are as in the table below:

PARAMETER	ESTIMATE	S. ERROR
geo_girl	0.411	0.1155
alg_girl	0.5892	0.1454
geo_cons	7.583	0.153
alg_cons	8.91	0.181
geo_base	1.526	0.09656
alg_base	1.843	0.1213
geo_base2	0.2309	0.06102
alg_base2	0.4747	0.07675
geo_base3	0.06106	0.03144
alg_base3	0.1585	0.03953

The estimated residual variance/correlation matrix at school level is

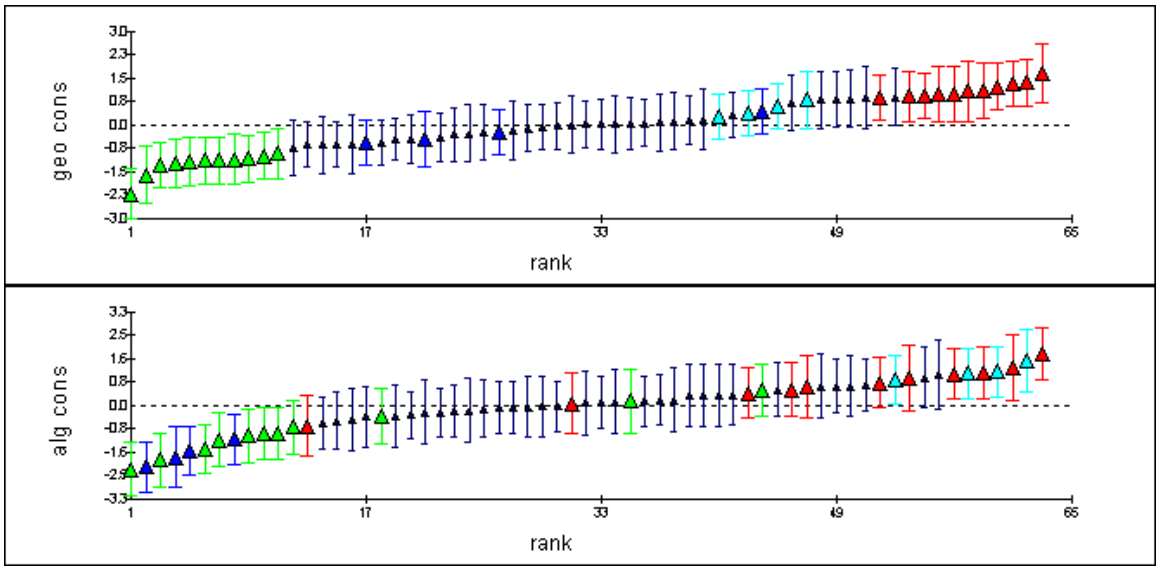
	Geo	*A*lg
Geo	0.845	
*A*lg	$r = .58$	1.066

,

which compares in an obvious way with that for Model 1. The residual variance/correlation matrix at student level is almost unchanged from Model 1. The full tabulation of the random part follows:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(geo_cons)	0.845	0.1872	1
σ_v(geo_cons,alg_cons)	0.5508	0.1703	0.58
σ_v^2(alg_cons)	1.066	0.2477	1
Student			
σ_u^2(geo_girl)	7.941	0.3095	1
σ_u(alg_girl,geo_girl)	1.816	0.2794	0.182
σ_u^2(alg_girl)	12.51	0.4878	1
σ_u^2(geo_boy)	8.449	0.3357	1
σ_u(alg_boy,geo_boy)	2.579	0.309	0.242
σ_u^2(alg_boy)	13.45	0.5359	1

School-level residuals against their ranks (Model 3):



School-level residual ranks (Model 3):

School order

School	Geom	Alg
1	49	42
2	42	37
3	4	12
4	7	44
5	40	48
6	44	8
7	60	63
8	22	26
9	20	49
10	38	41
11	8	18
12	52	52
13	37	19
14	17	5
15	2	35
16	21	4
17	50	40
18	25	17
20	55	43
21	41	60
22	34	39
23	26	2
24	62	59
25	39	45
26	57	47
27	32	14
28	29	30
29	11	3
30	63	61
31	56	13

Geometry order

School	Geom	Alg
30	63	61
24	62	59
59	61	57
7	60	63
61	59	31
56	58	54
26	57	47
31	56	13
20	55	43
57	54	46
55	53	34
12	52	52
53	51	56
17	50	40
1	49	42
51	48	32
64	47	62
33	46	50
62	45	53
6	44	8
60	43	58
2	42	37
21	41	60
5	40	48
25	39	45
10	38	41
13	37	19
52	36	51
38	35	25
22	34	39

Algebra order

School	Geom	Alg
7	60	63
64	47	62
30	63	61
21	41	60
24	62	59
60	43	58
59	61	57
53	51	56
48	23	55
56	58	54
62	45	53
12	52	52
52	36	51
33	46	50
9	20	49
5	40	48
26	57	47
57	54	46
25	39	45
4	7	44
20	55	43
1	49	42
10	38	41
17	50	40
22	34	39
40	16	38
2	42	37
45	18	36
15	2	35
55	53	34

32	24	24
33	46	50
34	14	28
35	3	6
36	1	1
37	27	21
38	35	25
39	10	10
40	16	38
41	12	29
42	6	11
43	13	27
44	19	20
45	18	36
46	5	7
47	28	33
48	23	55
49	9	9
50	30	23
51	48	32
52	36	51
53	51	56
54	33	15
55	53	34
56	58	54
57	54	46
58	15	22
59	61	57
60	43	58
61	59	31
62	45	53
63	31	16
64	47	62

54	33	15
27	32	14
63	31	16
50	30	23
28	29	30
47	28	33
37	27	21
23	26	2
18	25	17
32	24	24
48	23	55
8	22	26
16	21	4
9	20	49
44	19	20
45	18	36
14	17	5
40	16	38
58	15	22
34	14	28
43	13	27
41	12	29
29	11	3
39	10	10
49	9	9
11	8	18
4	7	44
42	6	11
46	5	7
3	4	12
35	3	6
15	2	35
36	1	1

47	28	33
51	48	32
61	59	31
28	29	30
41	12	29
34	14	28
43	13	27
8	22	26
38	35	25
32	24	24
50	30	23
58	15	22
37	27	21
44	19	20
13	37	19
11	8	18
18	25	17
63	31	16
54	33	15
27	32	14
31	56	13
3	4	12
42	6	11
39	10	10
49	9	9
6	44	8
46	5	7
35	3	6
14	17	5
16	21	4
29	11	3
23	26	2
36	1	1

Above the mean for geometry: 30, 24, 59, 7, 61, 56, 26, 31, 20, 57, 12
Above the mean for algebra: 7, 64, 30, 21, 24, 60, 59
Large differences between geometry and algebra rankings: 31, 4, 6, 15

Model 4 (as Model 2 but with no random effect of gender at school level)

The model for the fixed part is:

$$\begin{aligned}\text{predicted geometry score} &= 7.56 + 1.494base + 0.2271base^2 + 0.06258base^3 \\ &\quad + 0.4004girl + 0.5276 \leftarrow\!\!\!\!\!\!\% A\_C, \\ \text{predicted algebra score} &= 8.888 + 1.818base + 0.4713base^2 + 0.1596base^3 \\ &\quad + 0.5813girl + 0.464 \leftarrow\!\!\!\!\!\!\% A\_C,\end{aligned}$$

similar to Model 2. Standard errors are as in the table below:

PARAMETER	ESTIMATE	S. ERROR
geo_girl	0.4004	0.1153
alg_girl	0.5813	0.1454
geo_cons	7.56	0.1369
alg_cons	8.888	0.1703
geo_base	1.494	0.09689
alg_base	1.818	0.1222
geo_base2	0.2271	0.06085
alg_base2	0.4713	0.07666
geo_base3	0.06258	0.03135
alg_base3	0.1596	0.03948
geo_%_A_C	0.5276	0.105
alg_%_a_c	0.464	0.1297

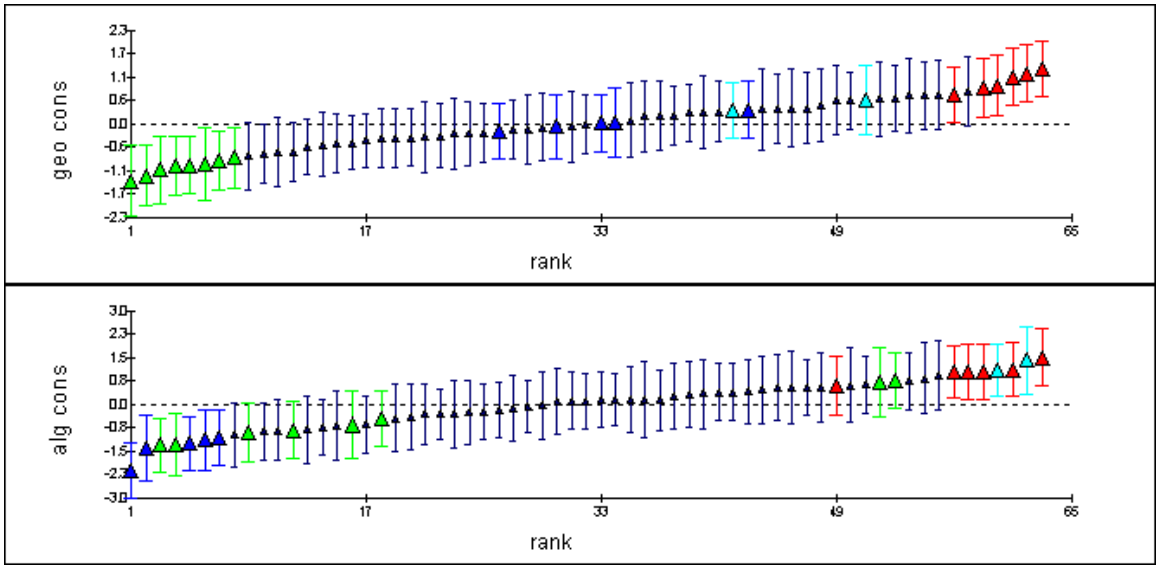
The estimated residual variance/correlation matrix at school level is:

	G	A
G	0.557	
A	$r = .43$	0.832

.

Compared to Model 2, it is relatively straightforward to detect correlation at school level between performance in Geometry and in Algebra. But this simpler model fails to detect significantly higher variance in boys' Algebra than in girls'. The full tabulation of the random part is:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(geo_cons)	0.5567	0.1355	1
σ_v(geo_cons,alg_cons)	0.2901	0.1259	0.426
σ_v^2(alg_cons)	0.8322	0.206	1
σ_u^2(geo_girl)	7.92	0.3084	1
σ_u(alg_girl,geo_girl)	1.799	0.2787	0.181
σ_u^2(alg_girl)	12.5	0.4872	1
σ_u^2(geo_boy)	8.46	0.3359	1
σ_u(alg_boy,geo_boy)	2.592	0.3092	0.243
σ_u^2(alg_boy)	13.46	0.5362	1



School-level residuals against their ranks (Model 4):

School-level residual rankings (Model 4):

School order		
School	Geom	Alg
1	31	19
2	39	33
3	2	12
4	7	53
5	48	56
6	43	5
7	59	63
8	15	26
9	18	54
10	28	35
11	14	37
12	61	58
13	23	15
14	33	7
15	6	52
16	34	2
17	58	39
18	29	24
20	50	32
21	19	51
22	36	44
23	26	1
24	62	59
25	32	41
26	45	30
27	25	10
28	20	25
29	30	6
30	44	36
31	55	8
32	40	40

Geometry order		
School	Geom	Alg
59	63	61
24	62	59
12	61	58
57	60	49
7	59	63
17	58	39
62	57	57
55	56	29
31	55	8
37	54	46
38	53	43
56	52	48
64	51	62
20	50	32
61	49	20
5	48	56
51	47	23
53	46	50
26	45	30
30	44	36
6	43	5
60	42	60
50	41	31
32	40	40
2	39	33
47	38	42
33	37	38
22	36	44
40	35	55
16	34	2
14	33	7

Algebra order		
School	Geom	Alg
7	59	63
64	51	62
59	63	61
60	42	60
24	62	59
12	61	58
62	57	57
5	48	56
40	35	55
9	18	54
4	7	53
15	6	52
21	19	51
53	46	50
57	60	49
56	52	48
52	27	47
37	54	46
48	9	45
22	36	44
38	53	43
47	38	42
25	32	41
32	40	40
17	58	39
33	37	38
11	14	37
30	44	36
10	28	35
45	13	34
2	39	33

33	37	38
34	11	27
35	5	3
36	3	4
37	54	46
38	53	43
39	12	14
40	35	55
41	1	16
42	10	17
43	4	18
44	17	22
45	13	34
46	8	9
47	38	42
48	9	45
49	24	21
50	41	31
51	47	23
52	27	47
53	46	50
54	22	11
55	56	29
56	52	48
57	60	49
58	16	28
59	63	61
60	42	60
61	49	20
62	57	57
63	21	13
64	51	62

25	32	41
1	31	19
29	30	6
18	29	24
10	28	35
52	27	47
23	26	1
27	25	10
49	24	21
13	23	15
54	22	11
63	21	13
28	20	25
21	19	51
9	18	54
44	17	22
58	16	28
8	15	26
11	14	37
45	13	34
39	12	14
34	11	27
42	10	17
48	9	45
46	8	9
4	7	53
15	6	52
35	5	3
43	4	18
36	3	4
3	2	12
41	1	16

20	50	32
50	41	31
26	45	30
55	56	29
58	16	28
34	11	27
8	15	26
28	20	25
18	29	24
51	47	23
44	17	22
49	24	21
61	49	20
1	31	19
43	4	18
42	10	17
41	1	16
13	23	15
39	12	14
63	21	13
3	2	12
54	22	11
27	25	10
46	8	9
31	55	8
14	33	7
29	30	6
6	43	5
36	3	4
35	5	3
16	34	2
23	26	1

Above average for geometry: 59, 24, 12, 57, 7, 62
Above average for algebra: 7, 64, 59, 60, 24, 12, 62
Large differences between algebra and geometry rankings: 31, 4, 15, 6

Large changes in rankings from Model 3 to Model 4

geometry		algebra	
sch	change	sch	change
21	down 22	30	down 25
30	down 19	1	down 23
1	down 18	26	down 17
37	up 27	37	up 25
40	up 19	11	up 19
29	up 19	38	up 18

This completes the modelling of constructive proof scores for the purpose of ranking schools. We now elaborate the model to explore other fixed effects.

2.2 A model for explanation

The other variables found to have a statistically significant effect on overall scores for constructive proof were:

textbook
VR score
choice of own approach to A3

Note that choice for best mark in A3 or G3 was *not* a statistically significant predictor for the corresponding constructive proof score. Of the three variables above I regard the latter two as 'correlated outcomes': it is unclear whether, for example, a high VR score can be said to predict or explain a high constructive proof score. It is just as likely that the 'arrow of prediction' goes the other way, or that there is no arrow of prediction at all between these two outcomes – and similarly for other pairs of proof-score outcomes. As to choice of own approach to A3, it turns out that those who choose the empirical options A or C are predicted to score more highly on constructive proof than those who choose B or D. Option B, however, would be the best choice. The prediction, therefore, is perverse, and probably arises from the influence of a few schools with high-scoring students who consistently follow this pattern.

Model 5 includes effects of textbook and VR score. For the reasons just given, care should be taken not to over-interpret the 'effect sizes' in the model.

We first give the distributions of Algebra VR score (algVR) and Geometry VR score (geoVR), together with their means and standard deviations:

Distribution of Algebra VR score

```
Each * = up to 14 cases
    Lower
    limit       N

       0.0000    396 : ****************************
        1.000    255 : ******************
        2.000    605 : ******************************************
        3.000    351 : *************************
        4.000    686 : ************************************************
        5.000    217 : ***************
        6.000    148 : ***********
        7.000    141 : ***********
```

Distribution of Geometry VR score

```
Each * = up to 13 cases
    Lower
    limit       N

       0.0000    407 : *****************************
        1.000    380 : ***************************
        2.000    644 : *********************************************
        3.000    388 : ****************************
        4.000    557 : ****************************************
        5.000    184 : *************
        6.000    149 : ***********
        7.000     90 : *******
```

Means and SD of Algebra and Geometry VR score

	N	Missing	Mean	s.d.
algVR	2799	0	2.9375	1.9205
geoVR	2799	0	2.6810	1.8539

For Model 5, we have used the cubic relationship with baseline score to produce two converted baseline scores for each student, one called *geoformula* which is appropriate for predicting Geometry score and one called *algformula* for predicting Algebra score. The relationship between the outcome score and the corresponding converted base line score is now linear (see equation on p24). The means and SD of the converted scores are:

	N	Missing	Mean	s.d.
geoformula	2799	136	0.17592	1.5235
algformula	2799	136	0.35792	2.0011

The distributions of the converted scores are as follows.

Distribution of *geoformula*

```
136 missing value(s)

Each * = up to 9 cases
    Lower
    limit       N

       -6.000      1 : *
       -5.500      2 : *
       -5.000      2 : *
       -4.500     11 : **
       -4.000     15 : **
       -3.500     25 : ***
       -3.000     30 : ****
       -2.500    112 : ************
       -2.000     93 : ***********
       -1.500    312 : ***********************************
       -1.000    221 : *************************
       -0.500    446 : ************************************************
        0        289 : ********************************
        0.500    260 : ****************************
        1.000    269 : *****************************
        1.500    226 : *************************
        2.000    160 : ******************
        2.500    131 : ***************
        3.000      0 :
        3.500     58 : *******
        4.000      0 :
```

Distribution of *algformula*

```
136 missing value(s)

Each * = up to 9 cases
    Lower
    limit       N

       -8.500      1 : *
       -8.000      0 :
       -7.500      2 : *
       -7.000      0 :
       -6.500      2 : *
       -6.000      0 :
       -5.500     11 : **
       -5.000     15 : **
       -4.500      0 :
       -4.000     25 : ***
       -3.500     30 : ****
       -3.000     48 : ******
       -2.500     64 : ********
       -2.000    240 : **************************
       -1.500    165 : ******************
       -1.000    431 : ************************************************
       -0.500    236 : *************************
        0        289 : ********************************
        0.500    260 : ****************************
        1.000      0 :
        1.500    269 : *****************************
        2.000    226 : *************************
        2.500      0 :
        3.000    160 : ******************
        3.500      0 :
        4.000    131 : ***************
        4.500      0 :
        5.000     58 : *******
        5.500      0 :
```

The converted scores are standardised to *z*-scores in Model 5, which now follows.

Model 5

Interest in this model centres on the fixed part prediction, and we do not discuss the random part. We defer to Model 7 discussion of the relative sizes of the fixed and random effects.

The models for Geometry and Algebra in the fixed part are:

$$\text{predicted geometry score} = 7.757 + 1.523\,geoformula + 0.3913\,girl + 0.514 \leftarrow \%A\_C + 0.2312\,geoVR,$$

$$\text{predicted algebra score} = 9.116 + 2.001\,a\lg formula + 0.5406\,girl + 0.3993 \leftarrow \%A\_C + 1.209\,text2 + 0.3931\,a\lg \mathrm{VR},$$

with standard errors as under:

PARAMETER	ESTIMATE	S. ERROR
geo_girl	0.3913	0.1265
alg_girl	0.5406	0.1685
geo_cons	7.757	0.1285
alg_cons	9.116	0.1815
geo_%_a_c	0.514	0.1028
alg_%_a_c	0.3993	0.1084
alg_text2	1.209	0.3213
alg_VR	0.3931	0.07301
geo_VR	0.2312	0.05791
geo_formula	1.523	0.06486
alg_formula	2.001	0.08075

The VR scores for Geometry and Algebra have been standardised so that their coefficients represent the additional score on constructive proof that is associated with an increase of 1 SD in the corresponding VR score of the student. The baseline formula scores and School %A*-C have also been standardised across all students. These conversions allow the effects of the different predictors on a given outcome (algebra or geometry) to be compared. The coefficient of *text2* estimates the additional score that is associated with using textbook 2 by contrast with all the other texts. This effect is not statistically significant for Geometry, but for Algebra it is approximately double the gender effect and three times the effect of a difference of 1 SD in the school's %A*-C. The geometry and algebra outcome scores have *not* been standardised, so this model should not be used to compare effect sizes *across* the two outcomes. See Model 7 for this.

3 Validity Rating scores for Geometry and Algebra

Validity rating scores are correlated with scores for constructive proof, and we model them together as a multivariate outcome. As with the models in the previous section, this allows us to study residual correlations at school and student level.

3.1 A model for comparing schools

Our model for comparing schools corresponds to Model 1, above. It includes as fixed predictors gender and standardised base line score only, and includes random effects of gender at school and student levels.

Model 6

The model for the fixed part is:

$$\begin{aligned}
\text{predicted geometry score} &= 7.579 + 1.539base + 0.2268base^2 + 0.06009base^3 + 0.4097girl,\\
\text{predicted algebra score} &= 8.933 + 1.875base + 0.4777base^2 + 0.158base^3 + 0.5423girl,\\
\text{predicted geometry VR score} &= 2.512 + 0.5125base + 0.1847base^2 + 0.04043base^3,\\
\text{predicted algebra VR score} &= 2.78 + 0.4673base + 0.1323base^2 + 0.04587base^3 + 0.1396girl,
\end{aligned}$$

with standard errors as in the table below.

PARAMETER	ESTIMATE	S. ERROR
geo_cons	7.579	0.1499
alg_cons	8.933	0.1976
geoVR_cons	2.512	0.04577
algVR_cons	2.78	0.07125
geo_girl	0.4097	0.1235
alg_girl	0.5423	0.1657
algVR_girl	0.1396	0.07368
geo_base	1.539	0.09608
alg_base	1.875	0.1206
geoVR_base	0.5125	0.05542
algVR_base	0.4673	0.05831
geo_base2	0.2268	0.06082
alg_base2	0.4777	0.07642
geoVR_base2	0.1847	0.03642
algVR_base2	0.1323	0.03758
geo_base3	0.06009	0.03133
alg_base3	0.158	0.03936
geoVR_base3	0.04043	0.01862
algVR_base3	0.04587	0.01929

There is no statistically non-significant gender effect on geometry VR score. As with constructive scores, the VR scores are associated with a cubic function of baseline score.

The residual variance/correlation matrix at school level is (variances on the diagonal, correlations elsewhere):

		Girls		*Boys*		*Girls*		*Boys*	
		Geo	*Alg*	*Geo*	*Alg*	*GeoVR*	*AlgVR*	*GeoVR*	*AlgVR*
Girls	*Geo*	.833							
	Alg	$r = .47$	.660						
Boys	*Geo*	$r = .92$	$r = .54$	.787					
	Alg	$r = .59$	$r = .88$	$r = .49$	1.445				
Girls	*GeoVR*	*ns*	*ns*	*ns*	*ns*	*ns*			
	AlgVR	$r = .77$	$r = .22$ *ns*	$r = .41$ *ns*	$r = .42$ *ns*	*ns*	.101		
Boys	*GeoVR*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	
	AlgVR	$r = .57$	$r = .54$	$r = .53$	$r = .60$	*ns*	$r = .72$	*ns*	.088

There is no statistically significant variation at school level in Geometry VR score, and hence no corresponding correlations. (In a further analysis, not shown, it was confirmed that there is no statistically significant school-level variation in Geometry VR score, even when the boys' and girls' scores are pooled.) We retain the relevant rows and columns in the above matrix to preserve symmetry. Two of the three non-significant correlations with girls' Algebra VR score are close to statistical significance and we retain all three. To remove these correlations would distort the estimation of the others.

At student level, the residual variance/correlation matrix is:

		Girls		*Boys*		*Girls*		*Boys*	
		Geo	*Alg*	*Geo*	*Alg*	*GeoVR*	*AlgVR*	*GeoVR*	*AlgVR*
Girls	*Geo*	7.92							
	Alg	$r = .19$	12.54						
Boys	*Geo*			8.42					
	Alg			$r = .25$	13.29				
Girls	*GeoVR*	$r = .11$	$r = .14$			2.89			
	AlgVR	$r = .13$	$r = .11$			$r = .29$	2.95		
Boys	*GeoVR*			$r = .09$	$r = .18$			3.31	
	AlgVR			$r = .04$ *ns*	$r = .12$			$r = .36$	3.50

Note the low correlations between scores in different sections for both girls and boys, particularly between performance in constructive proof and VR score, once other effects (including school effects) are allowed for.

Tabulations of the random part of Model 6, including standard errors, are given on the following pages.

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(geo_girl)	0.8331	0.2153	1
σ_v(alg_girl,geo_girl)	0.3465	0.1636	0.467
σ_v^2(alg_girl)	0.6601	0.2214	1
σ_v(geo_boy,geo_girl)	0.7481	0.1804	0.924
σ_v(geo_boy,alg_girl)	0.3884	0.1635	0.539
σ_v^2(geo_boy)	0.7868	0.2161	1
σ_v(alg_boy,geo_girl)	0.6431	0.2193	0.586
σ_v(alg_boy,alg_girl)	0.8556	0.2334	0.876
σ_v(alg_boy,geo_boy)	0.5171	0.2178	0.485
σ_v^2(alg_boy)	1.445	0.3787	1
σ_v(algVR_girl,geo_girl)	0.222	0.07315	0.766
σ_v(algVR_girl,alg_girl)	0.05771	0.06776	0.224
σ_v(algVR_girl,geo_boy)	0.1156	0.06843	0.411
σ_v(algVR_girl,alg_boy)	0.1601	0.09095	0.42
σ_v^2(algVR_girl)	0.1007	0.04062	1
σ_v(algVR_boy,geo_girl)	0.1543	0.07091	0.571
σ_v(algVR_boy,alg_girl)	0.1301	0.0711	0.541
σ_v(algVR_boy,geo_boy)	0.1385	0.06975	0.527
σ_v(algVR_boy,alg_boy)	0.2145	0.09404	0.602
σ_v(algVR_boy,algVR_girl)	0.06789	0.03087	0.722
σ_v^2(algVR_boy)	0.08782	0.04166	1

PARAMETER	ESTIMATE	S. ERROR	CORR.
Student			
σ_u^2(geo_girl)	7.924	0.311	1
σ_u(alg_girl,geo_girl)	1.887	0.2817	0.189
σ_u^2(alg_girl)	12.54	0.4924	1
σ_u^2(geo_boy)	8.421	0.3371	1
σ_u(alg_boy,geo_boy)	2.619	0.3093	0.248
σ_u^2(alg_boy)	13.29	0.5337	1
σ_u(geoVR_girl,geo_girl)	0.5433	0.1319	0.113
σ_u(geoVR_girl,alg_girl)	0.8654	0.1664	0.144
σ_u^2(geoVR_girl)	2.892	0.1109	1
σ_u(algVR_girl,geo_girl)	0.6091	0.1351	0.126
σ_u(algVR_girl,alg_girl)	0.6778	0.1698	0.111
σ_u(algVR_girl,geoVR_girl)	0.8442	0.08297	0.289
σ_u^2(algVR_girl)	2.952	0.1155	1
σ_u(geoVR_boy,geo_boy)	0.4733	0.1482	0.0897
σ_u(geoVR_boy,alg_boy)	1.174	0.1892	0.177
σ_u^2(geoVR_boy)	3.309	0.1296	1
σ_u(algVR_boy,geo_boy)	0.2275	0.1535	0.0419
σ_u(algVR_boy,alg_boy)	0.794	0.1945	0.116
σ_u(algVR_boy,geoVR_boy)	1.232	0.1006	0.362
σ_u^2(algVR_boy)	3.497	0.1394	1

Since there is no significant residual variation in Geometry VR score at school level, we give the residual ranks in Algebra VR score only.

School residual rankings for Algebra VR score (Model 6):

School order		
School	algVR_girl	algVR_boy
1	39	N/A
2	43	37
3	7	4
4	23	41
5	40	N/A
6	21	35
7	61	59
8	47	46
9	4	9
10	33	43
11	3	11
12	62	57
13	18	18
14	44	15
15	2	10
16	13	3
17	57	34
18	49	31
20	60	52
21	54	58
22	19	13
23	6	7
24	58	56
25	45	29
26	37	38
27	34	30
28	24	32
29	48	40
30	63	N/A
31	53	N/A
32	17	20
33	59	53
34	32	23
35	9	6
36	8	1
37	42	42
38	26	24
39	22	16
40	11	19
41	27	22
42	1	2
43	14	8
44	51	21
45	29	33
46	10	5
47	5	12

Ranked by girls' VR scores		
School	algVR_girl	algVR_boy
30	63	N/A
12	62	57
7	61	59
20	60	52
33	59	53
24	58	56
17	57	34
57	56	54
50	55	47
21	54	58
31	53	N/A
64	52	55
44	51	21
53	50	50
18	49	31
29	48	40
8	47	46
60	46	49
25	45	29
14	44	15
2	43	37
37	42	42
59	41	36
5	40	N/A
1	39	N/A
56	38	44
26	37	38
54	36	26
51	35	27
27	34	30
10	33	43
34	32	23
63	31	25
62	30	48
45	29	33
61	28	28
41	27	22
38	26	24
58	25	45
28	24	32
4	23	41
39	22	16
6	21	35
55	20	14
22	19	13
13	18	18

Ranked by boys' VR scores		
School	algVR_girl	algVR_boy
7	61	59
21	54	58
12	62	57
24	58	56
64	52	55
57	56	54
33	59	53
20	60	52
48	12	51
53	50	50
60	46	49
62	30	48
50	55	47
8	47	46
58	25	45
56	38	44
10	33	43
37	42	42
4	23	41
29	48	40
52	15	39
26	37	38
2	43	37
59	41	36
6	21	35
17	57	34
45	29	33
28	24	32
18	49	31
27	34	30
25	45	29
61	28	28
51	35	27
54	36	26
63	31	25
38	26	24
34	32	23
41	27	22
44	51	21
32	17	20
40	11	19
13	18	18
49	16	17
39	22	16
14	44	15
55	20	14

48	12	51
49	16	17
50	55	47
51	35	27
52	15	39
53	50	50
54	36	26
55	20	14
56	38	44
57	56	54
58	25	45
59	41	36
60	46	49
61	28	28
62	30	48
63	31	25
64	52	55

32	17	20
49	16	17
52	15	39
43	14	8
16	13	3
48	12	51
40	11	19
46	10	5
35	9	6
36	8	1
3	7	4
23	6	7
47	5	12
9	4	9
11	3	11
15	2	10
42	1	2

22	19	13
47	5	12
11	3	11
15	2	10
9	4	9
43	14	8
23	6	7
35	9	6
46	10	5
3	7	4
16	13	3
42	1	2
36	8	1

Above the mean for girls' Algebra VR: 30, 12
Above the mean for boys' Algebra VR: 7, 21
Below the mean for girls' Algebra VR: 42, 15, 11, 9, 47, 23
Below the mean for boys' Algebra VR: 36, 42, 16, 3, 46, 35

3.2 Models for explanation

Our purpose is to show all the statistically significant effects and to compare effect sizes. For this, we normalise the outcome variables, so that they are on comparable scales. The raw-score scales for the outcome scores have the following distributions.

Distribution of Geometry constructive scores

```
Each * = up to 6 cases
    Lower
    limit       N

       0.0000     36 : ******
        1.000     53 : *********
        2.000    110 : ******************
        3.000    148 : ************************
        4.000    212 : ***********************************
        5.000    219 : ************************************
        6.000    264 : *********************************************
        7.000    291 : ************************************************
        8.000    288 : ************************************************
        9.000    277 : **********************************************
        10.00    255 : ******************************************
        11.00    228 : **************************************
        12.00    160 : ***************************
        13.00    103 : ******************
        14.00    136 : ************************
        15.00     19 : ****
```

Distribution of Algebra constructive scores

```
11 missing value(s)

Each * = up to 6 cases
    Lower
    limit       N

       0.0000     60 : **********
        1.000     44 : ********
        2.000     53 : *********
        3.000    134 : ***********************
        4.000    118 : ********************
        5.000    148 : *************************
        6.000    216 : ************************************
        7.000    182 : ******************************
        8.000    241 : ****************************************
        9.000    269 : *********************************************
        10.00    202 : **********************************
        11.00    220 : ************************************
        12.00    245 : ****************************************
        13.00    166 : ****************************
        14.00    142 : ************************
        15.00    118 : ********************
        16.00     77 : *************
        17.00     59 : **********
        18.00     41 : *******
        19.00     24 : ****
        20.00     17 : ***
        21.00      7 : **
        22.00      5 : *
```

Distribution of Geometry VR scores

```
Each * = up to 13 cases
    Lower
    limit       N

       0.0000    407 : ********************************
        1.000    380 : ******************************
        2.000    644 : **************************************************
        3.000    388 : ******************************
        4.000    557 : *******************************************
        5.000    184 : **************
        6.000    149 : ************
        7.000     90 : *******
```

Distribution of Algebra VR scores

```
Each * = up to 14 cases
    Lower
    limit       N

       0.0000    396 : ******************************
        1.000    255 : ********************
        2.000    605 : ***********************************************
        3.000    351 : ***************************
        4.000    686 : ****************************************************
        5.000    217 : *****************
        6.000    148 : ***********
        7.000    141 : ***********
```

Means and standard deviations

	N	Missing	Mean	s.d.
geo_total	2799	0	7.8955	3.5215
alg_total	2799	11	9.5203	4.4073
geoVR_total	2799	0	2.6810	1.8539
algVR_total	2799	0	2.9375	1.9205

We have already described the standardisation of school %A*-C. For baseline scores, we generated for each outcome (Geometry constructive score, Algebra constructive score, Geometry VR score, Algebra VR score) a converted baseline score for each student so that the relationship between outcome and converted baseline score was linear. These converted scores are called, respectively, *geoformula*, *algformula*, *geoVRformula*, *algVRformula*.

Distribution of *geoformula*

```
136 missing value(s)

Each * = up to 11 cases
    Lower
    limit       N

       -1.800      1 : *
       -1.600      2 : *
       -1.400      2 : *
       -1.200     26 : ***
       -1.000     25 : ***
      -0.8000     78 : ********
      -0.6000    304 : ****************************
      -0.4000    386 : ************************************
      -0.2000    446 : *****************************************
  -1.665e-016    289 : ***************************
       0.2000    529 : **************************************************
       0.4000    226 : *********************
       0.6000    160 : ***************
       0.8000    131 : ************
        1.000     58 : ******
        1.200      0 :
```

Distribution of *algformula*

```
136 missing value(s)

Each * = up to 11 cases
    Lower
    limit       N

       -2.200      1 : *
       -2.000      2 : *
       -1.800      0 :
       -1.600      2 : *
       -1.400     11 : *
       -1.200     15 : **
       -1.000     55 : *****
      -0.8000     48 : *****
      -0.6000    157 : ***************
      -0.4000    533 : ************************************************
      -0.2000    446 : ****************************************
  -3.886e-016    289 : ***************************
       0.2000    529 : ************************************************
       0.4000    226 : *********************
       0.6000    160 : ***************
       0.8000    131 : ************
        1.000      0 :
        1.200     58 : ******
        1.400      0 :
```

Distribution of *geoVRformula*

```
136 missing value(s)

Each * = up to 11 cases
    Lower
    limit       N

      -0.7000      1 : *
      -0.6000      2 : *
      -0.5000      2 : *
      -0.4000     51 : *****
      -0.3000    235 : **********************
      -0.2000    533 : ************************************************
      -0.1000    446 : ****************************************
  -1.388e-016    289 : ***************************
       0.1000    260 : ************************
       0.2000    269 : *************************
       0.3000    226 : *********************
       0.4000    160 : ***************
       0.5000      0 :
       0.6000    131 : ************
       0.7000     58 : ******
       0.8000      0 :
```

Distribution of *algVRformula*

```
136 missing value(s)

Each * = up to 11 cases
    Lower
    limit       N

       -1.300      1 : *
       -1.200      0 :
       -1.100      2 : *
       -1.000      0 :
      -0.9000      2 : *
      -0.8000     11 : *
      -0.7000     15 : **
      -0.6000     25 : ***
      -0.5000     30 : ***
      -0.4000    112 : ***********
      -0.3000    240 : **********************
      -0.2000    386 : ************************************
      -0.1000    446 : *****************************************
  -2.776e-017    289 : ***************************
       0.1000    529 : ************************************************
       0.2000    226 : *********************
       0.3000      0 :
       0.4000    160 : ***************
       0.5000    131 : ************
       0.6000      0 :
       0.7000     58 : ******
       0.8000      0 :
```

Means and standard deviations

	N	Missing	Mean	s.d.
geoformula	2799	136	0.061045	0.44710
algformula	2799	136	0.084050	0.47800
geoVRformula	2799	136	0.079165	0.25656
algVRformula	2799	136	0.048575	0.26624

Model 7

The model for the fixed effects is:

$$\text{normalised geometry score} = -0.039 + 0.446\,geoformula + 0.109\,girl + 0.146 \leftarrow\!\!\!\%A\_C,$$

$$\text{normalised algebra score} = -0.086 + 0.476\,a\lg formula + 0.123\,girl + 0.098 \leftarrow\!\!\!\%A\_C + 0.247\,text2,$$

$$\text{normalised geometry VR score} = -0.011 + 0.256\,geoVRformula + 0.071 \leftarrow\!\!\!\%A\_C + 0.121\,text2,$$

$$\text{normalised algebra VR score} = -0.017 + 0.257\,a\lg VRformula + 0.068\,girl + 0.063 \leftarrow\!\!\!\%A\_C,$$

where *geoformula*, etc., have been standardised to *z*-scores.

The standard errors are as under:

PARAMETER	ESTIMATE	S. ERROR
geo_cons	-0.03868	0.03582
alg_cons	-0.08595	0.04067
geoVR_cons	-0.01087	0.01864
algVR_cons	-0.01738	0.03078
geo_girl	0.1087	0.03458
alg_girl	0.1227	0.03748
algVR_girl	0.06773	0.03574
geo_formula	0.4456	0.01768
alg_formula	0.4761	0.01764
geoVR_formula	0.2558	0.01794
algVR_formula	0.257	0.0185
alg_text2	0.2469	0.06745
geoVR_text2	0.121	0.04559
geo_%_a_c	0.1458	0.02837
alg_%_a_c	0.0975	0.02396
geoVR_%_a_c	0.0708	0.01797
algVR_%_a_c	0.06272	0.02343

We can see that the effect of baseline score on geometry and algebra constructive scores is similar, as it is on VR scores for the two subjects. (In fact, a more precise picture requires that we take account of the distribution of the standardised converted scores in each case. See below, where the above observation is confirmed for all except the very lowest scoring students.)

There is, as we have seen, no gender effect on Geometry VR score. The girls are predicted higher scores on all 3 other outcomes, with the greatest advantage apparently in Algebra constructive proof. These apparent differences are not statistically significant, however.

Indeed, some care is needed when attempting to contrast effects. There is a significant difference between the effects of the baseline score on constructive and VR scores, and between the effects of school GCSE %A*-C on Geometry constructive and VR scores. But other apparent differences between corresponding effects are not significant.

Textbook 2 has a positive effect (in contrast to other texts) on not only Algebra constructive proof but also Geometry VR score.

Changing the scale of the outcome variables changes the residual variance, and including the fixed effects of textbook and %A*-C further reduces school-level variation. Some correlations, in particular, are poorly estimated and as a result fail to reach statistical significance. We retain them in the model, for comparison with Model 6. Like Model 6, Model 7 is unable to estimate residual variances and covariances at school level when these relate to Geometry VR scores. The estimated residual variance/correlation matrix at school level for Model 7 is:

		Girls		*Boys*		*Girls*		*Boys*	
		Geo	*Alg*	*Geo*	*Alg*	*GeoVR*	*AlgVR*	*GeoVR*	*AlgVR*
Girls	*Geo*	.042							
	Alg	$r = .23\,ns$	.014						
Boys	*Geo*	$r = .89$	$r = .31\,ns$	.045					
	Alg	$r = .45$	$r = .91$	$r = .31\,ns$	.060				
Girls	*GeoVR*	*ns*	*ns*	*ns*	*ns*	*ns*			
	AlgVR	$r = .75$	$r = .01\,ns$	$r = .33\,ns$	$r = .30\,ns$	*ns*	.023		
Boys	*GeoVR*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	
	AlgVR	$r = .55\,ns$	$r = .58\,ns$	$r = .52\,ns$	$r = .52\,ns$	*ns*	$r = .72$	*ns*	.020

At student level, it is:

		Girls		*Boys*		*Girls*		*Boys*	
		Geo	*Alg*	*Geo*	*Alg*	*GeoVR*	*AlgVR*	*GeoVR*	*AlgVR*
Girls	*Geo*	.62							
	Alg	$r = .19$	.63						
Boys	*Geo*			.67					
	Alg			$r = .24$	.68				
Girls	*GeoVR*	$r = .12$	$r = .14$			.74			
	AlgVR	$r = .13$	$r = .11$			$r = .29$	.69		
Boys	*GeoVR*			$r = .09$	$r = .18$			.84	
	AlgVR			$r = .04\,ns$	$r = .12$			$r = .36$	.83

,

where the correlations are virtually identical to those in Model 6.

The random part of Model 7 is tabulated on the following pages.

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(geo_girl)	0.0421	0.01271	1
σ_v(alg_girl,geo_girl)	0.005546	0.007112	0.226
σ_v^2(alg_girl)	0.01436	0.00768	1
σ_v(geo_boy,geo_girl)	0.0388	0.01072	0.892
σ_v(geo_boy,alg_girl)	0.007878	0.007487	0.31
σ_v^2(geo_boy)	0.04497	0.01396	1
σ_v(alg_boy,geo_girl)	0.02253	0.01085	0.449
σ_v(alg_boy,alg_girl)	0.0266	0.008824	0.907
σ_v(alg_boy,geo_boy)	0.01593	0.01123	0.307
σ_v^2(alg_boy)	0.05991	0.01682	1
σ_v(algVR_girl,geo_girl)	0.02341	0.008441	0.753
σ_v(algVR_girl,alg_girl)	0.0001519	0.006021	0.00837
σ_v(algVR_girl,geo_boy)	0.01044	0.008316	0.325
σ_v(algVR_girl,alg_boy)	0.01094	0.009136	0.295
σ_v^2(algVR_girl)	0.02297	0.009421	1
σ_v(algVR_boy,geo_girl)	0.01614	0.008268	0.553
σ_v(algVR_boy,alg_girl)	0.009932	0.006365	0.583
σ_v(algVR_boy,geo_boy)	0.01566	0.008526	0.52
σ_v(algVR_boy,alg_boy)	0.01813	0.009437	0.521
σ_v(algVR_boy,algVR_girl)	0.01547	0.00718	0.718
σ_v^2(algVR_boy)	0.0202	0.009744	1

PARAMETER	ESTIMATE	S. ERROR	CORR.
Student			
σ_u^2(geo_girl)	0.6236	0.02447	1
σ_u(alg_girl,geo_girl)	0.1173	0.0177	0.187
σ_u^2(alg_girl)	0.6301	0.02473	1
σ_u^2(geo_boy)	0.6674	0.02672	1
σ_u(alg_boy,geo_boy)	0.1638	0.01965	0.243
σ_u^2(alg_boy)	0.6785	0.02724	1
σ_u(geoVR_girl,geo_girl)	0.08127	0.01866	0.12
σ_u(geoVR_girl,alg_girl)	0.09569	0.01875	0.141
σ_u^2(geoVR_girl)	0.7361	0.02824	1
σ_u(algVR_girl,geo_girl)	0.08209	0.01835	0.125
σ_u(algVR_girl,alg_girl)	0.07074	0.0184	0.107
σ_u(algVR_girl,geoVR_girl)	0.2033	0.02025	0.285
σ_u^2(algVR_girl)	0.6918	0.02708	1
σ_u(geoVR_boy,geo_boy)	0.06685	0.02095	0.0895
σ_u(geoVR_boy,alg_boy)	0.1341	0.02148	0.178
σ_u^2(geoVR_boy)	0.8352	0.03271	1
σ_u(algVR_boy,geo_boy)	0.03313	0.02099	0.0446
σ_u(algVR_boy,alg_boy)	0.08743	0.02135	0.117
σ_u(algVR_boy,geoVR_boy)	0.2977	0.02453	0.359
σ_u^2(algVR_boy)	0.825	0.03289	1

We now give a list of school-level residuals, not for the purpose of ranking, but in order to illustrate their sizes when compared to the fixed effects.

School-level residuals (Model 7):

School	geo_girl	geo_boy	alg_girl	alg_boy	algVR_girl	algVR_boy
1	-0.037	N/A	-0.031	N/A	-0.053	N/A
2	0.119	0.074	0.022	0.101	0.058	0.037
3	-0.375	-0.375	-0.086	-0.202	-0.203	-0.224
4	-0.232	-0.245	0.155	0.264	-0.032	0.048
5	0.125	N/A	0.133	N/A	0.076	N/A
6	0.042	0.194	-0.087	-0.309	-0.033	0.038
7	0.254	0.213	0.104	0.240	0.145	0.173
8	-0.035	-0.129	-0.105	-0.186	0.119	0.050
9	-0.164	-0.093	0.016	-0.017	-0.224	-0.158
10	-0.045	-0.023	0.053	0.073	-0.002	0.037
11	-0.183	-0.135	0.027	-0.053	-0.206	-0.058
12	0.364	0.291	0.108	0.299	0.215	0.222
13	-0.080	-0.041	-0.064	-0.167	-0.089	-0.084
14	0.112	-0.019	-0.129	-0.124	0.144	-0.048
15	-0.321	-0.290	0.134	0.183	-0.236	-0.098
16	-0.054	0.014	-0.234	-0.473	-0.067	-0.174
17	0.223	0.150	-0.010	0.098	0.135	0.046
18	0.065	-0.063	-0.036	0.027	0.118	0.024
20	0.145	0.166	0.040	0.101	0.116	0.091
21	-0.017	-0.124	0.138	0.284	0.074	0.133
22	-0.012	0.000	0.007	0.047	-0.046	-0.064
23	-0.107	0.061	-0.246	-0.645	-0.185	-0.151
24	0.342	0.341	0.150	0.343	0.136	0.172
25	0.025	-0.040	0.045	0.175	0.045	-0.007
26	0.092	0.115	-0.075	-0.135	0.014	-0.025
27	-0.055	-0.043	-0.084	-0.188	0.015	-0.020
28	-0.126	-0.062	0.013	-0.037	-0.048	-0.012
29	0.100	0.035	-0.112	-0.206	0.195	0.103
30	0.156	N/A	0.021	N/A	0.149	N/A
31	0.207	N/A	-0.112	N/A	0.102	N/A
32	0.058	0.080	0.035	0.056	-0.014	0.031
33	0.095	0.036	0.033	0.112	0.108	0.089
34	-0.133	-0.217	-0.015	0.016	0.011	-0.036
35	-0.319	-0.265	-0.141	-0.349	-0.142	-0.174
36	-0.278	-0.343	-0.155	-0.289	-0.086	-0.185
37	0.255	0.198	0.057	0.154	0.153	0.171
38	0.132	0.177	0.048	0.097	0.047	0.068
39	-0.142	-0.182	-0.047	-0.082	-0.013	-0.069
40	-0.029	0.053	0.137	0.204	-0.096	0.029
41	-0.335	-0.400	-0.067	-0.134	-0.065	-0.109
42	-0.276	-0.148	-0.028	-0.154	-0.267	-0.220
43	-0.271	-0.332	-0.070	-0.105	-0.157	-0.194
44	0.000	-0.160	-0.074	-0.013	0.104	-0.031
45	-0.131	-0.161	0.043	0.080	0.006	0.010
46	-0.224	-0.271	-0.139	-0.276	-0.127	-0.166
47	-0.077	0.091	0.090	0.078	-0.206	-0.086
48	-0.240	-0.130	0.178	0.168	-0.192	0.050
49	-0.055	-0.030	0.013	-0.007	-0.004	0.007
50	0.150	0.091	0.037	0.135	0.166	0.116
51	0.098	0.117	-0.005	0.019	-0.008	-0.037
52	-0.081	0.003	0.022	-0.051	-0.079	0.006

53	0.138	0.095	0.098	0.249	0.053	0.062
54	-0.065	-0.067	-0.095	-0.203	0.007	-0.029
55	0.105	0.156	-0.022	-0.013	-0.067	-0.089
56	0.172	0.183	0.105	0.233	0.000	0.035
57	0.258	0.276	0.087	0.181	0.125	0.168
58	-0.111	-0.104	-0.111	-0.325	0.029	0.086
59	0.293	0.347	0.022	0.102	0.087	0.031
60	0.059	0.100	0.032	0.050	0.095	0.110
61	0.092	0.169	-0.049	-0.115	-0.033	-0.045
62	0.157	0.206	0.160	0.291	0.011	0.124
63	-0.031	-0.053	-0.043	-0.071	-0.007	-0.047
64	0.207	0.120	0.109	0.305	0.130	0.119

Comparative effect sizes, from Model 7

To compare the relative sizes of the different effects we tabulate below the estimated contribution to each normalised score of school residuals, base line score, and school GCSE %A*-C at the top, the 85th centile, the 15th centile, and the bottom of the sample distribution in each case. The final table shows the effects of gender and textbook 2.

School residuals:

	geo_girl	geo_boy	alg_girl	alg_boy	geoVR_girl	geoVR_boy	algVR_girl	algVR_boy
Top	0.364	0.347	0.178	0.343	0	0	0.215	0.222
85th	0.172	0.177	0.108	0.233	0	0	0.130	0.116
15th	-0.224	-0.217	-0.105	-0.202	0	0	-0.142	-0.109
Bottom	-0.375	-0.400	-0.246	-0.645	0	0	-0.267	-0.224

Baseline score:

	Geo	Alg	Geo VR	Alg VR
Top	1.016	1.173	0.709	0.640
85th	0.419	0.446	0.248	0.240
15th	-0.558	-0.552	-0.279	-0.295
Bottom	-1.762	-2.254	-0.689	-1.223

School % A*-C:

	Geo	Alg	Geo VR	Alg VR
Top	0.480	0.320	0.233	0.206
85th	0.151	0.101	0.073	0.065
15th	-0.146	-0.097	-0.070	-0.063
Bottom	-0.343	-0.229	-0.167	-0.148

Gender and Textbook:

	Geo	Alg	Geo VR	Alg VR
Girl	0.109	0.123	0	0.0768
Text 2	0	0.247	0.121	0

These effects are additive. Simply looking at the ranges in each column in the first and third tables we can see that school effects due to unmeasured characteristics (the residuals) cover approximately the same range as the effects of schools' % GCSE A*-C (except for Geometry VR, where there were no detectable school residual effects). The significantly higher school-level variance in boys' than in girls' algebra scores is reflected in the columns headed 'alg_girl' and 'alg_boy' in the first table, where the range for boys is seen to be about twice the range for girls. The effect on algebra scores of using textbook 2, compared to all other texts, exceeds the residual effects of all but a small number of schools, and is similar to the maximum effects (positive or negative) of the school's % GCSE A*-C, when compared to the average. By far the most important contribution comes from the student's relative ranking on the baseline score.

We may further illustrate these effects by considering four specific schools, two high-ranking, one middle-ranking, and one low (in that order):

	Residuals							
School	geo_girl	geo_boy	alg_girl	alg_boy	algVR_girl	algVR_boy	% A*-C	Textbook
7	0.25	0.21	0.10	0.24	0.15	0.17	61%	2
59	0.29	0.35	0.02	0.10	0.09	0.03	48%	2
34	-0.13	-0.22	-0.02	0.02	0.01	-0.04	56%	5
23	-0.11	0.06	-0.25	-0.65	-0.19	-0.15	52%	5

On the following page we show the predicted normalised outcome scores, together with the effects on them of

- student baseline score
- student gender
- school residual
- school % A*-C
- textbook

for girls and boys at the 15th, 50th, and 85th centiles of their distribution of baseline scores. The girls, on average, scored about 1 raw-score point less than the boys on the baseline test, and this shows up in the 15th and 50th centiles. The 85th centile is 19 points for both girls and boys.

Alongside the predicted normalised score (headed 'norm') is given the equivalent raw score obtained by linear interpolation. The normalising transformation is in each case approximately linear over the outcome range involved.

As an example of the use of the table, consider a girl in school 7 who is at the 50th centile of the baseline score distribution for girls. To find her predicted algebra score we go to the section headed 'Algebra', and the column headed 'raw' under School 7. The predicted score is 11.0. This arises by linear interpolation from a normal score of 0.33, and this latter score is made up of six components: the intercept (not shown in this table, but in the description of the fixed part of Model 7), equal to -0.09 (2 d.p.); the contribution from the baseline formula, equal to -0.11(see the column headed 'Formula'); the gender effect (see the column headed 'Girl'), equal to 0.12; the girl residual for the school, equal to 0.10; the school %A*-C effect, equal to 0.07; and, finally, the textbook effect, equal to 0.25. The same girl would be predicted to score -0.33 on the normalised scale, translating to 8.1 raw-score points, if she were in school 29. This school does not use textbook 2, and has a lower %A*-C and a negative girl residual. These three differences account for the difference in the predicted score and this difference is the same (apart from rounding error) for a girl at any point on the baseline score scale. The difference for boys is greater, since the difference between the boy residuals for the two schools is greater. *The table is illustrative only. Errors of estimation accumulate when effects are aggregated, and the estimated differences shown for these particular schools may not exceed the accumulated error in their estimation.*

Normalised outcome scores, and effects on them, together with interpolated raw scores

Geometry				School 7		School 59		School 34		School 29	
	Baseline	Formula	Girl	norm	*raw*	norm	*raw*	norm	*raw*	norm	*raw*
Girl, 15th centile	10	-0.56	0.11	-0.14	*7.3*	-0.23	*7.0*	-0.57	*5.7*	-0.59	*5.6*
Boy, 15th centile	11	-0.47		-0.20	*7.1*	-0.19	*7.1*	-0.68	*5.3*	-0.44	*6.2*
Girl, 50th centile	15	-0.09	0.11	0.33	*9.1*	0.24	*8.8*	-0.10	*7.5*	-0.12	*7.4*
Boy, 50th centile	16	0.02		0.29	*9.0*	0.30	*9.0*	-0.19	*7.1*	0.05	*8.1*
Girl, 85th centile	19	0.43	0.11	0.85	*11.1*	0.76	*10.8*	0.42	*9.5*	0.40	*9.4*
Boy, 85th centile	19	0.43		0.70	*10.6*	0.72	*10.6*	0.22	*8.7*	0.47	*9.7*
Girl residual				0.25		0.29		-0.13		-0.11	
Boy residual				0.21		0.35		-0.22		0.06	
%A*-C effect				0.10		-0.03		0.05		0.01	
Textbook effect				0		0		0		0	

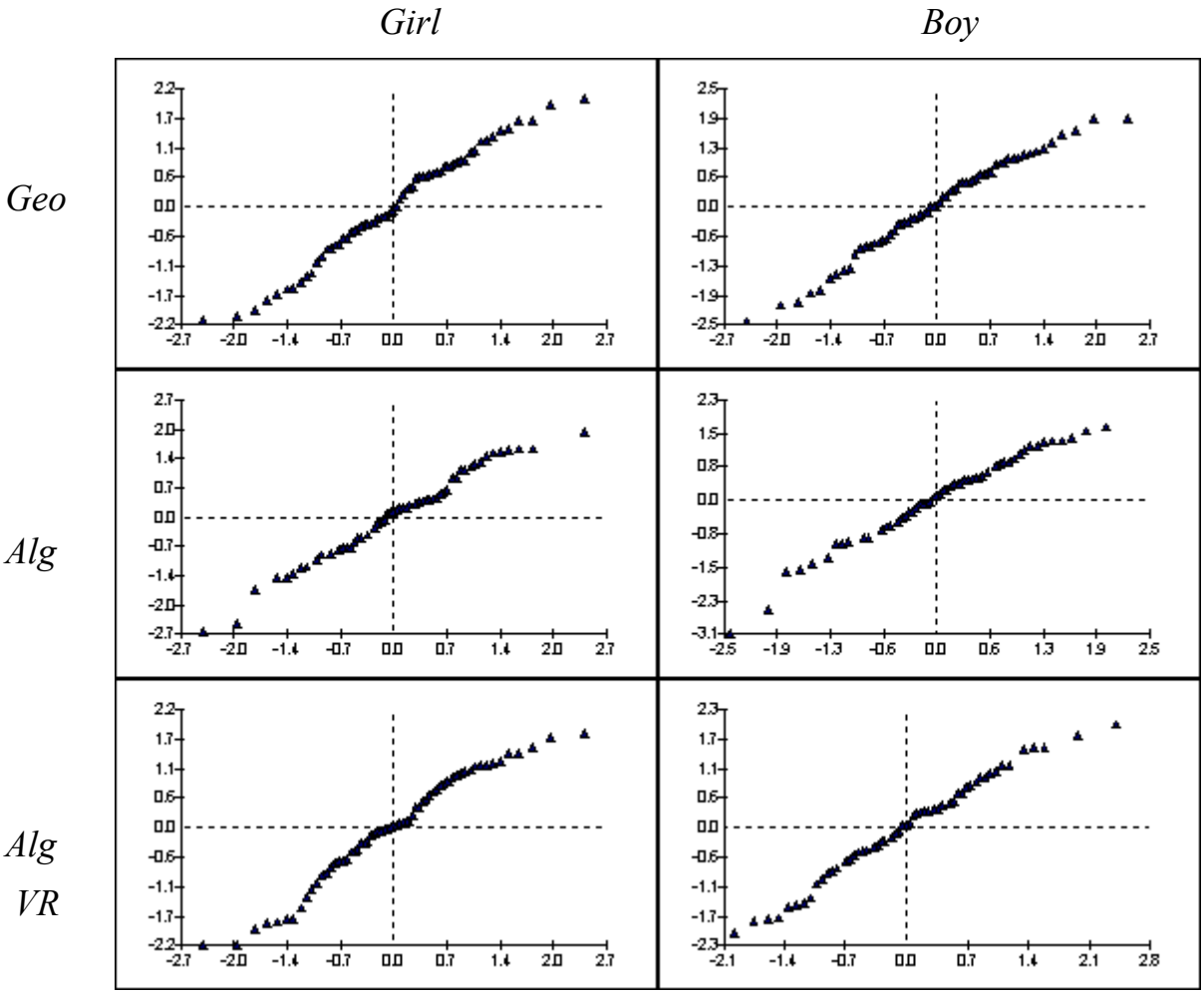
Algebra				School 7		School 59		School 34		School 29	
	Baseline	Formula	Girl	norm	*raw*	norm	*raw*	norm	*raw*	norm	*raw*
Girl, 15th centile	10	-0.55	0.12	-0.11	*9.1*	-0.27	*8.3*	-0.51	*7.3*	-0.77	*6.1*
Boy, 15th centile	11	-0.46		0.00	*9.5*	-0.23	*8.5*	-0.50	*7.3*	-1.20	*4.1*
Girl, 50th centile	15	-0.11	0.12	0.33	*11.0*	0.16	*10.3*	-0.07	*9.2*	-0.33	*8.1*
Boy, 50th centile	16	-0.01		0.45	*11.6*	0.23	*10.6*	-0.05	*9.3*	-0.74	*6.2*
Girl, 85th centile	19	0.45	0.12	0.89	*13.5*	0.73	*12.8*	0.49	*11.7*	0.23	*10.6*
Boy, 85th centile	19	0.45		0.91	*13.6*	0.68	*12.6*	0.41	*11.4*	-0.29	*8.2*
Girl residual				0.10		0.02		-0.02		-0.25	
Boy residual				0.24		0.10		0.02		-0.65	
%A*-C effect				0.07		-0.02		0.04		0.01	
Textbook effect				0.25		0.25		0		0	

Geometry VR				School 7		School 59		School 34		School 29	
	Baseline	Formula	Girl	norm	*raw*	norm	*raw*	norm	*raw*	norm	*raw*
Girl, 15th centile	10	-0.28		-0.12	*2.4*	-0.18	*2.3*	-0.26	*2.1*	-0.28	*2.0*
Boy, 15th centile	11	-0.25		-0.09	*2.4*	-0.16	*2.3*	-0.24	*2.1*	-0.26	*2.1*
Girl, 50th centile	15	-0.10		0.06	*2.8*	0.00	*2.6*	-0.08	*2.5*	-0.10	*2.4*
Boy, 50th centile	16	-0.03		0.13	*2.9*	0.06	*2.8*	-0.02	*2.6*	-0.04	*2.5*
Girl, 85th centile	19	0.25		0.41	*3.5*	0.34	*3.3*	0.26	*3.2*	0.24	*3.1*
Boy, 85th centile	19	0.25		0.41	*3.5*	0.34	*3.3*	0.26	*3.2*	0.24	*3.1*
Girl residual				0		0		0		0	
Boy residual				0		0		0		0	
%A*-C effect				0.05		-0.01		0.03		0.01	
Textbook effect				0.12		0.12		0		0	

Algebra VR				School 7		School 59		School 34		School 29	
	Baseline	Formula	Girl	norm	*raw*	norm	*raw*	norm	*raw*	norm	*raw*
Girl, 15th centile	10	-0.30	0.07	-0.06	*2.8*	-0.17	*2.5*	-0.22	*2.4*	-0.44	*1.9*
Boy, 15th centile	11	-0.25		-0.06	*2.8*	-0.25	*2.3*	-0.29	*2.3*	-0.42	*2.0*
Girl, 50th centile	15	-0.06	0.07	0.17	*3.3*	0.06	*3.0*	0.01	*2.9*	-0.20	*2.4*
Boy, 50th centile	16	0.00		0.19	*3.3*	-0.01	*2.9*	-0.05	*2.8*	-0.17	*2.5*
Girl, 85th centile	19	0.24	0.07	0.48	*3.9*	0.36	*3.7*	0.32	*3.6*	0.10	*3.1*
Boy, 85th centile	19	0.24		0.43	*3.8*	0.24	*3.4*	0.20	*3.3*	0.07	*3.0*
Girl residual				0.15		0.09		0.01		-0.19	
Boy residual				0.17		0.03		-0.04		-0.15	
%A*-C effect				0.04		-0.01		0.02		0.01	
Textbook effect				0		0		0		0	

Diagnostic residuals

The following are the plots of standardised diagnostic school-level residuals against their normal scores.



These are adequate.

Model 8

Finally, seeing proof as general (as indicated by a response to L1b coded 30) has an effect on all scores. Here the base category, obviously, is those students whose response to L1b was coded other than 30. The fixed-part model is:

$$\begin{aligned}
\text{normalised geometry score} &= -0.095 + 0.431 geoformula + 0.108 girl + 0.148 \leftarrow\!\!\!\!\!\!\not\%\, A_C \\
&\quad + 0.121 \leftarrow\!\!\!\!\!\!\not L1b30, \\
\text{normalised algebra score} &= -0.348 + 0.401 a\lg formula + 0.093 girl + 0.087 \leftarrow\!\!\!\!\!\!\not\%\, A_C + 0.223 text2 \\
&\quad + 0.608 \leftarrow\!\!\!\!\!\!\not L1b30, \\
\text{normalised geometry VR score} &= -0.100 + 0.231 geoVRformula + 0.067 \leftarrow\!\!\!\!\!\!\not\%\, A_C + 0.109 text2 \\
&\quad + 0.195 \leftarrow\!\!\!\!\!\!\not L1b30, \\
\text{normalised algebra VR score} &= -0.098 + 0.232 a\lg VRformula + 0.060 girl + 0.064 \leftarrow\!\!\!\!\!\!\not\%\, A_C \\
&\quad + 0.181 \leftarrow\!\!\!\!\!\!\not L1b30,
\end{aligned}$$

with standard errors as under:

PARAMETER	ESTIMATE	S. ERROR
geo_cons	-0.09452	0.03876
alg_cons	-0.3477	0.04082
geoVR_cons	-0.1002	0.02492
algVR_cons	-0.09841	0.03514
geo_girl	0.1079	0.03397
alg_girl	0.09268	0.03631
algVR_girl	0.05971	0.03679
geo_%_a_c	0.1481	0.02754
alg_%_a_c	0.08744	0.02318
geoVR_%_a_c	0.06744	0.01797
algVR_%_a_c	0.06365	0.02297
alg_text2	0.2226	0.06679
geoVR_text2	0.1091	0.04567
geo_formula	0.4307	0.01819
alg_formula	0.4008	0.01712
geoVR_formula	0.2311	0.01849
algVR_formula	0.2316	0.01911
geo_l1b30	**0.1211**	**0.0327**
alg_l1b30	**0.608**	**0.03092**
geoVR_l1b30	**0.1946**	**0.0355**
algVR_l1b30	**0.1807**	**0.03516**

The large effect on algebra constructive scores is due to the fact that a response to question L1b coded 30 adds 2 to the algebra score.

The random part of this model is of no especial interest.

4 Scores on individual constructive proof questions

4.1 Models for explanation

Model 9

We begin with a model of Geometry scores. We model scores on G1, G2a, G2b, and G4 as a multivariate outcome. The distributions of these separate outcome scores are as follows.

Distribution of G1 score

```
Each * = up to 26 cases
    Lower
    limit       N

       0.0000   1300 : **************************************************
       0.5000      0 :
        1.000    171 : *******
        1.500      0 :
        2.000    214 : *********
        2.500    571 : **********************
        3.000    543 : *********************
        3.500      0 :
```

Distribution of G2a score

```
Each * = up to 23 cases
    Lower
    limit       N

       0.0000   1101 : ***********************************************
       0.5000      0 :
        1.000    476 : ********************
        1.500      0 :
        2.000    117 : ******
        2.500    788 : *********************************
        3.000    317 : **************
        3.500      0 :
```

Distribution of G2b score

```
Each * = up to 23 cases
    Lower
    limit       N

       0.0000    453 : ********************
        1.000    493 : **********************
        2.000    737 : *********************************
        3.000   1116 : **************************************************
```

Distribution of G4 score

```
Each * = up to 14 cases
    Lower
    limit       N

       0.0000    336 : ************************
        1.000    178 : *************
        2.000    492 : ************************************
        3.000    290 : *********************
        4.000    694 : **************************************************
        5.000    164 : ************
        6.000    645 : ***********************************************
```

The distributions of the G1 and G2a scores, in particular, are problematic for modelling as continuous variables.

Baseline score, gender, school %A*-C, and Geometry VR score are found to have effects on at least three of the four outcomes. Textbook 2 has an effect on the score for G1 but not the other

questions. 'Proof as general' (code 30 on L1b) has an effect on the score for G4, but not the others. There are also random effects of gender at school and student levels.

We find that, for G1 and G2a, a quadratic function of standardised baseline score is required. For G1 the function is

$$g1formula \ldots 0.415base + 0.063base^2$$

and for G2a it is

$$g2aformula \ldots 0.287base + 0.050base^2.$$

For G2b we require a cubic function:

$$g2bformula \ldots 0.157base - 0.018base^2 + 0.022base^3,$$

while for G4 the standardised baseline score, which for this purpose we shall call *g*4*base*, is sufficient by itself.

The new functions have the following distributions.

Distribution of *g*1*formula*

```
136 missing value(s)

Each * = up to 7 cases
    Lower
    limit       N

      -0.7000     86 : *************
      -0.6000    112 : *****************
      -0.5000     93 : **************
      -0.4000    312 : ************************************************
      -0.3000    221 : **********************************
      -0.2000    210 : ********************************
      -0.1000    236 : ************************************
  -1.388e-016    289 : ********************************************
       0.1000    260 : ****************************************
       0.2000      0 :
       0.3000    269 : *****************************************
       0.4000    226 : **********************************
       0.5000      0 :
       0.6000    160 : ************************
       0.7000    131 : ********************
       0.8000      0 :
       0.9000     58 : *********
        1.000      0 :
```

Distribution of *g*2*aformula*

```
136 missing value(s)

Each * = up to 9 cases
    Lower
    limit       N

      -0.5000     53 : ******
      -0.4000    238 : ***************************
      -0.3000    312 : ************************************
      -0.2000    221 : *************************
      -0.1000    446 : ***************************************************
  -2.776e-017    289 : *********************************
       0.1000    260 : ******************************
       0.2000    269 : *******************************
       0.3000    226 : **************************
       0.4000    160 : ******************
       0.5000    131 : ***************
       0.6000     58 : *******
       0.7000      0 :
```

Distribution of *g2bformula*

```
136 missing value(s)

Each * = up to 11 cases
    Lower
    limit       N

       -2.100      1 : *
       -2.000      0 :
       -1.900      0 :
       -1.800      2 : *
       -1.700      0 :
       -1.600      0 :
       -1.500      2 : *
       -1.400      0 :
       -1.300     11 : *
       -1.200      0 :
       -1.100     15 : **
       -1.000      0 :
      -0.9000     25 : ***
      -0.8000      0 :
      -0.7000     30 : ***
      -0.6000     48 : *****
      -0.5000     64 : ******
      -0.4000     93 : *********
      -0.3000    147 : **************
      -0.2000    386 : ************************************
      -0.1000    446 : *****************************************
   6.384e-016    549 : ***************************************************
       0.1000    495 : ***********************************************
       0.2000    291 : ***************************
       0.3000     58 : ******
       0.4000      0 :
```

Means and standard deviations

	N	Missing	Mean	s.d.
g1formula	2799	136	0.062581	0.39132
g2aformula	2799	136	0.049702	0.26908
g2bformula	2799	136	-0.029239	0.24344

We standardise these functions to *z*-scores across the students and can then fit them to a linear relationship with the corresponding outcome scores. The resulting model for the fixed part is:

$$\begin{aligned}
\text{predicted G1 score} &= 1.366 + 0.390 \leftarrow\!\!\!\!\!\!\text{g1}\,formula \quad - 0.156\,girl + 0.092 \leftarrow\!\!\!\!\!\!\%A_C + 0.051\,VR \\
&\quad + 0.198\,text2, \\
\text{predicted G2a score} &= 1.307 + 0.268 \leftarrow\!\!\!\!\!\!\text{g2}\,aformula \quad\quad + 0.081 \leftarrow\!\!\!\!\!\!\%A_C + 0.046\,VR, \\
\text{predicted G2b score} &= 1.833 + 0.243 \leftarrow\!\!\!\!\!\!\text{g2}\,bformula + 0.144\,girl + 0.098 \leftarrow\!\!\!\!\!\!\%A_C, \\
\text{predicted G4 score} &= 3.134 + 0.649 \leftarrow\!\!\!\!\!\!\text{g4}\,base \quad\quad + 0.376\,girl + 0.237 \leftarrow\!\!\!\!\!\!\%A_C + 0.163\,VR \\
&\quad + 0.174 \leftarrow\!\!\!\!\!\!\text{L1}b30.
\end{aligned}$$

This model is suitable for comparing effect sizes within question, but not across questions, since the outcome scales are different.

VR is the standardised Geometry VR score, which is associated with improved scores on all Geometry constructive questions except G2b. Note the first instance of a negative gender effect for girls – for the G1 score. Textbook 2 is found to have a positive effect on scores for G1, but not the other Geometry constructive scores. Seeing proof as general, as indicated by the response 30 to question L1b, has a positive effect on the G4 score.

The standard errors of the fixed parameter estimates are given in the table on the following page.

PARAMETER	ESTIMATE	S. ERROR
g1_cons	1.366	0.0444
g2a_cons	1.307	0.02249
g2b_cons	1.833	0.02952
g4_cons	3.134	0.0743
g1_formula	0.3898	0.02649
g2a_formula	0.2676	0.02452
g2b_formula	0.2433	0.02205
g4_base	0.6473	0.03997
g1_girl	-0.1558	0.0486
g2b_girl	0.1444	0.04143
g4_girl	0.3756	0.07791
g1_%_a_c	0.09163	0.03335
g2a_%_a_c	0.08083	0.02379
g2b_%_a_c	0.09823	0.02166
g4_%_a_c	0.2374	0.05648
g1_vr	0.05072	0.02451
g2a_vr	0.04591	0.0236
g4_vr	0.1631	0.03568
g1_text2	0.1976	0.08792
g4_l1b30	0.1736	0.07043

Turning to the random part of Model 9, we find that at school level there is no statistically significant residual variation in scores for G2a or G2b. For the other two questions, schools vary slightly more for girls than for boys, though not significantly so. The estimated residual variance/correlation matrix at school level is:

		Girl		*Boy*	
		*G*1	*G*4	*G*1	*G*4
Girl	*G*1	0.038			
	*G*4	$r = .47\,ns$	0.183		
Boy	*G*1	$r = .92$	$r = .23\,ns$	0.037	
	*G*4	$r = .49\,ns$	$r = .76$	$r = .75$	0.124

where '*ns*' indicates poor estimation of these correlations. They are retained in the model, because there is *prima facie* reason to expect them and to remove them distorts the estimation of the remaining correlations. For example, there is high residual correlation at school level between girls' and boys' G1 scores and significant correlation between boys' G1 scores and boys' G4 scores. Therefore, to remove the correlation between girls' G1 and boys' G4 scores because it fails to reach statistical significance, effectively constraining this correlation to be zero, depresses the estimate of correlation between boys' G1 and G4 scores. In general, schools vary slightly more for girls than for boys, though not statistically significantly so.

The student-level residual variance/correlation matrix is estimated to be:

		Girl				*Boy*			
		*G*1	*G2a*	*G2b*	*G*4	*G*1	*G2a*	*G2b*	*G*4
	*G*1	1.39							
Girl	*G2a*	$r = .04$	1.36						
	G2b	$r = .09$	$r = .08$	1.13					
	*G*4	$r = .05$	$r = .05$	$r = .06$	2.84				
	*G*1					1.42			
Boy	*G2a*					$r = .07$	1.30		
	G2b					$r = .11$	$r = .12$	1.12	
	*G*4					$r = .02\,ns$	$r = .08$	$r = .09$	3.03

Girls' and boys' variances for the same questions are similar, and the correlations between their residual scores on different questions are low. In other words, once adjustment has been made for the fixed effects and the school residual effects, individual students who perform, for example, better than expected on G1 are not especially likely to perform better than expected on any of the other Geometry constructive questions. The full tabulation follows.

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(g1_girl)	0.03805	0.01842	1
σ_v(g4_girl, g1_girl)	0.03922	0.02355	0.471
σ_v^2(g4_girl)	0.1826	0.05688	1
σ_v(g1_boy, g1_girl)	0.03467	0.01411	0.919
σ_v(g1_boy, g4_girl)	0.01927	0.02365	0.233
σ_v^2(g1_boy)	0.03739	0.01872	1
σ_v(g4_boy, g1_girl)	0.03346	0.02194	0.486
σ_v(g4_boy, g4_girl)	0.1143	0.04047	0.759
σ_v(g4_boy, g1_boy)	0.05099	0.02247	0.748
σ_v^2(g4_boy)	0.1243	0.04857	1
Student			
σ_u^2(g1_girl)	1.393	0.05494	1
σ_u(g2a_girl, g1_girl)	0.04913	0.03781	0.0357
σ_u^2(g2a_girl)	1.362	0.0525	1
σ_u(g2b_girl, g1_girl)	0.1165	0.03452	0.093
σ_u(g2b_girl, g2a_girl)	0.1007	0.03389	0.0813
σ_u^2(g2b_girl)	1.127	0.04346	1
σ_u(g4_girl, g1_girl)	0.09487	0.05551	0.0477
σ_u(g4_girl, g2a_girl)	0.1075	0.05426	0.0547
σ_u(g4_girl, g2b_girl)	0.1103	0.04939	0.0617
σ_u^2(g4_girl)	2.835	0.1119	1
σ_u^2(g1_boy)	1.422	0.05727	1
σ_u(g2a_boy, g1_boy)	0.1009	0.03819	0.0742
σ_u^2(g2a_boy)	1.299	0.05122	1
σ_u(g2b_boy, g1_boy)	0.1446	0.03561	0.114
σ_u(g2b_boy, g2a_boy)	0.1449	0.03389	0.12
σ_u^2(g2b_boy)	1.121	0.04421	1
σ_u(g4_boy, g1_boy)	0.05147	0.05914	0.0248
σ_u(g4_boy, g2a_boy)	0.1501	0.05586	0.0756
σ_u(g4_boy, g2b_boy)	0.1747	0.05198	0.0948
σ_u^2(g4_boy)	3.03	0.1221	1

Model 10

This is the model for Algebra constructive scores (including Logic) that corresponds to Model 9 for Geometry. The distributions of the relevant scores are as follows.

Distribution of A1 score

```
Each * = up to 28 cases
    Lower
    limit       N

       0.0000   1395 : **************************************************
        1.000    131 : *****
        2.000     90 : ****
        3.000   1183 : ******************************************
```

Distribution of A2 score

```
Each * = up to 21 cases
    Lower
    limit       N

       0.0000   1047 : **************************************************
       0.5000      0 :
        1.000    110 : ******
        1.500      0 :
        2.000    191 : **********
        2.500    510 : *************************
        3.000    941 : ***********************************************
        3.500      0 :
```

Distribution of A4 score

```
2 missing value(s)

Each * = up to 22 cases
    Lower
    limit       N

       0.0000    232 : ***********
       0.5000    212 : **********
        1.000    363 : ******************
        1.500      0 :
        2.000      0 :
        2.500      0 :
        3.000    194 : *********
        3.500    648 : ********************************
        4.000   1070 : **************************************************
        4.500      0 :
        5.000      0 :
        5.500      2 : *
        6.000      5 : *
        6.500     24 : **
        7.000     47 : ***
        7.500      0 :
```

Distribution of L1 score

```
10 missing value(s)

Each * = up to 11 cases
    Lower
    limit       N

       0.0000    528 : ************************************************
       0.5000      0 :
        1.000     44 : ****
        1.500      0 :
        2.000    538 : *************************************************
        2.500     41 : ****
        3.000    223 : ********************
        3.500      6 : *
        4.000    316 : ****************************
        4.500     62 : ******
        5.000    320 : ****************************
        5.500     23 : ***
        6.000    288 : *************************
        6.500     32 : ***
        7.000    153 : **************
        7.500     17 : **
        8.000     94 : *********
        8.500     16 : **
        9.000     61 : ******
        9.500      3 : *
        10.00     24 : ***
        10.50      0 :
```

These distributions are far from ideal for modelling as continuous responses.

The relationship between A1 score and baseline score is quadratic:

$$a1formula \ldots 0.387base + 0.095base^2,$$

while that between L1 score and baseline score is cubic:

$$L1formula \ldots 0.628base + 0.221base^2 + 0.079base^3.$$

The relationships for A2 and A3 are linear, and for clarity we shall use $a2base$ and $a3base$ when describing these relationships.

The functions $a1formula$ and $L1formula$ are standardised across students to be able to compare effect sizes within these questions. The initial distributions of the functions are as follows.

Distribution of *a1formula*

```
136 missing value(s)

Each * = up to 9 cases
    Lower
    limit       N

      -0.4000    433 : ************************************************
      -0.3000    167 : ******************
      -0.2000    434 : *************************************************
      -0.1000    236 : *************************
  -2.776e-017    289 : ********************************
       0.1000    260 : ****************************
       0.2000      0 :
       0.3000    269 : *****************************
       0.4000    226 : *************************
       0.5000      0 :
       0.6000    160 : ******************
       0.7000    131 : ***************
       0.8000      0 :
       0.9000     58 : *******
        1.000      0 :
```

Distribution of *L1formula*

```
136 missing value(s)

Each * = up to 9 cases
    Lower
    limit       N

       -3.600      1 : *
       -3.400      0 :
       -3.200      0 :
       -3.000      2 : *
       -2.800      0 :
       -2.600      2 : *
       -2.400      0 :
       -2.200      0 :
       -2.000     11 : **
       -1.800     15 : **
       -1.600      0 :
       -1.400     25 : ***
       -1.200     30 : ****
       -1.000     48 : ******
      -0.8000    157 : ******************
      -0.6000    312 : ************************************
      -0.4000    221 : *************************
      -0.2000    446 : ***************************************************
  -3.886e-016    289 : *********************************
       0.2000    260 : ******************************
       0.4000    269 : *******************************
       0.6000      0 :
       0.8000    226 : **************************
        1.000      0 :
        1.200    160 : ******************
        1.400      0 :
        1.600    131 : ***************
        1.800      0 :
        2.000      0 :
        2.200     58 : *******
        2.400      0 :
```

Means and standard deviations

	N	Missing	Mean	s.d.
a1formula	2799	136	0.095335	0.36001
L1formula	2799	136	0.18131	0.77477

The fixed part of Model 10 is:

$$\begin{aligned}
\text{predicted A1 score} &= 1.289 + 0.360 \leftarrow a1formula + 0.079 \leftarrow \%A_C \\
&\quad + 0.618 text2, \\
\text{predicted A2 score} &= 1.457 + 0.441 \leftarrow a2base + 0.182 girl + 0.110 \leftarrow \%A_C \\
&\quad + 0.165 text2 + 0.164 L1b30, \\
\text{predicted A4 score} &= 2.694 + 0.410 \leftarrow a4base + 0.252 girl + 0.085 \leftarrow \%A_C \\
&\quad + 0.161 VR + 0.252 L1b30, \\
\text{predicted L1 score} &= 3.621 + 0.775 \leftarrow L1formula + 0.143 \leftarrow \%A_C \\
&\quad + 0.192 VR.
\end{aligned}$$

In these equations, *VR* indicates standardised validity rating score for Algebra. This has an effect on scores for A4 and L1. The gender effect is statistically detectable only for A2 and A4 scores. Textbook 2 has an effect on scores for A1 and A2. And 'proof-as-general' (choice 30 for L1b) has an effect on A2 and A4. (We excluded it from the model for L1, obviously.) The standard errors are detailed on the next page:

PARAMETER	ESTIMATE	S. ERROR
a1_cons	1.289	0.03896
a2_cons	1.457	0.04724
a4_cons	2.703	0.0604
L1_cons	3.621	0.05509
a1_formula	0.36	0.02838
a2_base	0.4405	0.02666
a4_base	0.4099	0.03374
L1_formula	0.7747	0.05222
a2_girl	0.1816	0.05417
a4_girl	0.2516	0.06534
a1_%a_c	0.07949	0.03558
a2_%a_c	0.1099	0.03183
a4_%a_c	0.08462	0.04227
L1_%a_c	0.1433	0.05613
a4_VR	0.1607	0.03039
L1_VR	0.1922	0.0496
a1_text2	0.6179	0.1004
a2_text2	0.1647	0.08095
a2_l1b30	0.1637	0.04887
a4_l1b30	0.252	0.05978

The estimated residual school-level variance/correlation matrix is:

		Girl				*Boy*			
		*A*1	*A*2	*A*4	*L*1	*A*1	*A*2	*A*4	*L*1
Girl	*A*1	.040							
	*A*2	*r* = .79	.052						
	*A*4	*ns*	*ns*	.078					
	*L*1	*ns*	*r* = .95*	*ns*	.076*				
Boy	*A*1	*ns*	*ns*	*ns*	*ns*	.128			
	*A*2	*ns*	*ns*	*ns*	*ns*	*r* = .73	.047		
	*A*4	*ns*	*ns*	*r* = .66	*ns*	*ns*	*ns*	.070	
	*L*1	*ns*	*ns*	*ns*	*ns*	*ns*	*r* = .84	*ns*	.129

Here, so many of the correlations are poorly estimated that we have removed them from the model, concluding that there is no justification for retaining them. In particular, the only statistically significant correlation between girls' and boys' residuals for the same question is that for A4. There is no reason to retain the others. The other school-level residual correlations that can be estimated satisfactorily are between A1 and A2, and between A2 and L1, for boys and girls separately. Note that the variance of the girls' L1 scores is poorly estimated (marked with an asterisk). This casts doubt on the very high estimate of correlation between girls' residual scores on A2 and L1 (also marked with an asterisk). Schools are statistically significantly more variable in their boys' than in their girls' scores on A1, but there is no statistically significant difference in the residual variation of girls' and boys' scores on the other questions.

The estimated residual student-level variance/correlation matrix is:

		Girl				*Boy*			
		*A*1	*A*2	*A*4	*L*1	*A*1	*A*2	*A*4	*L*1
	*A*1	1.79							
Girl	*A*2	$r = .10$	1.49						
	*A*4	*ns*	$r = .03$ *ns*	2.04					
	*L*1	*ns*	*ns*	*ns*	5.54				
	*A*1	0	0	0	0	1.75			
Boy	*A*2	0	0	0	0	$r = .14$	1.42		
	*A*4	0	0	0	0	*ns*	$r = .07$	2.15	
	*L*1	0	0	0	0	*ns*	*ns*	*ns*	5.89

As with the school-level correlations, where student-level correlations for the same pairs of questions were non-significant for both girls and boys we have omitted them. The only statistically significant student-level residual correlations found are those shown, and we have include the non-significant girls' correlation for A2 and A4 for consistency. All of these correlations are low. Student-level residual variances for boys and girls for the same questions are similar.

The full tabulation for the random parameters is:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(a1_girl)	0.03978	0.02213	1
σ_v(a2_girl,a1_girl)	0.03571	0.01581	0.787
σ_v^2(a2_girl)	0.05174	0.02153	1
σ_v^2(a4_girl)	0.07842	0.0314	1
σ_v(L1_girl,a2_girl)	0.06002	0.02479	0.954
σ_v^2(L1_girl)	0.07643	0.05979	1
σ_v^2(a1_boy)	0.1278	0.03912	1
σ_v(a2_boy,a1_boy)	0.05633	0.02103	0.725
σ_v^2(a2_boy)	0.04724	0.02028	1
σ_v(a4_boy,a4_girl)	0.04871	0.02346	0.659
σ_v^2(a4_boy)	0.06957	0.03127	1
σ_v(L1_boy,a2_boy)	0.06582	0.02613	0.843
σ_v^2(L1_boy)	0.1291	0.07389	1

PARAMETER	ESTIMATE	S. ERROR	CORR.
Student			
σ_u^2(a1_girl)	1.787	0.0705	1
σ_u(a2_girl,a1_girl)	0.167	0.04567	0.102
σ_u^2(a2_girl)	1.485	0.05859	1
σ_u(a4_girl,a2_girl)	0.0442	0.04786	0.0254
σ_u^2(a4_girl)	2.041	0.08055	1
σ_u^2(L1_girl)	5.543	0.2188	1
σ_u^2(a1_boy)	1.747	0.07045	1
σ_u(a2_boy,a1_boy)	0.2264	0.04522	0.144
σ_u^2(a2_boy)	1.418	0.05711	1
σ_u(a4_boy,a2_boy)	0.117	0.04881	0.0671
σ_u^2(a4_boy)	2.147	0.0866	1
σ_u^2(L1_boy)	5.894	0.2379	1

5 Choice for own approach in question G3

There are four possibilities for the student to choose from when stating their own approach to question G3, namely, A, B, C, and D. These have been coded 1, 2, 3, and 4, respectively. Other responses (91 etc) are treated as missing (116 out of 2799). Response A (considered 'empirical') is used as the base category and a multinomial model with logit link (Healy and Hoyles, June 1999, App. 5) (H&H) is used to estimate effects on the odds in favour of making one of the other choices, compared to the base. The estimates themselves are expressed as logarithms of the odds ratios in each case: thus a negative parameter estimate for a given predictor means that the corresponding choice becomes less likely (in comparison to the 'empirical' choice) as the value of that variable increases. The effect concerned is indicated by the prefix '2_', '3_', or '4_' in front of the variable name, thus prefix '2_' indicates an effect on the probability of making choice B, '3_' of making choice C, and '4_' of choice D.

Choice of predictors has been guided to some extent by the earlier survey. Some predictors available in that dataset (for example, student views of the role of proof, teacher approaches to the teaching of proof) are not available in this; and the categories of response are not strictly comparable; but, in the more general sense that H&H found that certain variables had an effect on student choice, an attempt has been made to replicate those findings where possible.

Initially, both school- and class-level variation were included. No significant variation, however, was found at class level within school, so this level is omitted from the analyses presented here. Thus, level 3 is school, with levels 1 and 2 reserved as usual for specifying the multinomial distribution (see Goldstein, 1995, pp104 *et seq*). Estimation is by penalized quasi-likelihood, with 2nd-order correction (*PQL-2*), allowing extra-multinomial variation. (The latter was found to be small throughout). The strategy for screening variables was to start with a model of the intercepts, and then test each variable in turn by itself, using a Wald test. This test provides an approximate chi-squared statistic that can be used to test the statistical significance of the effect of a variable on the three choices coded 2, 3, and 4, separately and jointly.

The way the fixed part in these models is interpreted is as follows. The sum of the probabilities of the choices A, B, C, and D is, of course, 1. Suppose the probability of making the 'empirical' choice (A, the base, coded 1) is *k*. Then the probability of making choice B (coded 2) is $k\exp(\ell_2)$, that of making choice C (3) $k\exp(\ell_3)$, and D (4) $k\exp(\ell_4)$, where we estimate $\ell_2, \ell_3,$ and ℓ_4. These are thus logarithmic relative probabilities for the choices B, C, and D, by comparison with the probability of choice A. A negative value of ℓ indicates a lower probability than that of the empirical choice, and a positive value a higher probability. *k*, of course, will be such that $k\left(1+\exp(\ell_2)+\exp(\ell_3)+\exp(\ell_4)\right)=1$. Each additive term in an expression for ℓ is converted into a multiplicative factor by the exponential function, so the coefficients are sometimes known as 'log odds ratios'. Discussion of the fixed effects centres on these logarithms, rather than on the resultant probabilities, because the sign of the logarithm immediately indicates the direction of the effect, and the Wald test indicates its statistical significance.

The fixed-part prediction in the basic model with just the intercept terms may be written:
$\text{predicted } \ell_2 = -0.09373,$
$\text{predicted } \ell_3 = -1.323,$
$\text{predicted } \ell_4 = -1.216.$

Choices 3 and 4 are significantly less likely than the empirical choice. The full tabulation is:

```
PARAMETER              ESTIMATE     S. ERROR
2_cons                 -0.09373       0.0639
3_cons                   -1.323      0.09478
4_cons                   -1.216      0.08496
```

There is significant school-level variation. The residual variance/correlation matrix at school level is estimated to be:

	B	*C*	*D*
B	.155		
C	$r = .89$	.318	
D	$r = .63$	$r = .65$	.228.

The full tabulation of the random part is:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(2_cons)	0.1553	0.04534	1
σ_v(3_cons,2_cons)	0.1967	0.05147	0.885
σ_v^2(3_cons)	0.3179	0.09931	1
σ_v(4_cons,2_cons)	0.1188	0.04367	0.631
σ_v(4_cons,3_cons)	0.1758	0.06571	0.653
σ_v^2(4_cons)	0.2282	0.07987	1
Extra-multinomial	0.9641	0.01537	

The 'extra-multinomial' estimate of 0.9641 indicates that the estimated residual variation in the response proportions once the other effects have been included is slightly less than would be predicted by the multinomial distribution. (The value 0.9641 is the ratio of this estimate to the expected variation.) This small 'under-dispersion' may indicate the existence of other effects, not included in the model, tending to induce further correlations between student responses.

The variables found to be significant predictors on their own were:

School level

area
minutes of maths per week (*lotsofmaths*, meaning more than 3 hours per week)

Student level

gender
baseline score
choice for best mark, where this was own choice
validity score for own choice
view of explanatory power of own choice
total score on G1, G2, and G4 (constructive proof)
total VR score for G3
proof-as-general (response coded 30 on L1b)

Not all of these were significant for all choices and not all remained significant with other variables in the model (see below). Since we cannot interpret an area effect, this is omitted.

Other variables tested but found to be non-significant were:

School level

school gender
textbook
GCSE syllabus
school admin (VA, etc.)
%A*-C
existence of maths club
school 11-18 status

Class level

teacher experience
teacher maths degree
teacher maths cert
teacher own choice for G3

We now give the estimates for three models, Model 11, which is suitable for comparing schools, and Models 12 and 13, which are alternative ways to explain what is going on at the student level.

5.1 A model for comparing schools

Model 11

This model contains terms for intercept, baseline score, and gender only. (Recall that school GCSE %A*-C is not a significant predictor.) The fixed-part prediction is:

$$\text{predicted } \ell_2 = -0.17\,base - 0.31\,girl,$$
$$\text{predicted } \ell_3 = -0.98 - 0.90\,girl,$$
$$\text{predicted } \ell_4 = -0.87 - 0.83\,girl,$$

with standard errors as under:

PARAMETER	ESTIMATE	S. ERROR
3_cons	-0.9816	0.109
4_cons	-0.8744	0.1029
2_base	-0.1682	0.04096
2_girl	-0.3132	0.07018
3_girl	-0.9041	0.1327
4_girl	-0.8307	0.1268

Baseline score has no statistically significant effect on the probability of choosing options C or D, compared with option A. In other words, all students are less likely to choose C or D than A, and the difference in likelihood does not depend on baseline score. Girls are still less likely to make these choices than boys. There is no statistically significant difference in the intercepts for options A and B. Boys scoring around the overall mean on the baseline test are as likely to choose option B as option A. Girls are less likely to choose option B than boys with the same baseline score. The likelihood of choosing option B decreases as baseline score increases.

The school-level residual variance/correlation matrix is:

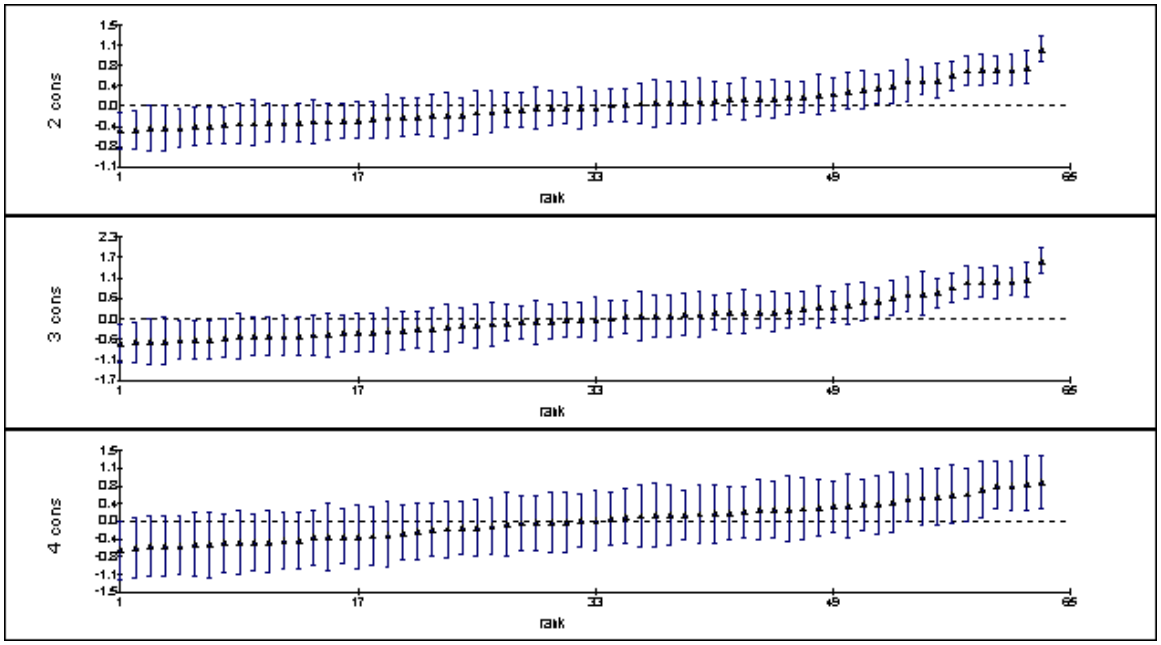
	B	*C*	*D*
B	.146		
C	$r = .97$	.349	
D	$r = .59$	$r = .57$	.244

.

Note the very high correlation at school level between choices B and C. This means that a school with a higher than predicted proportion of students opting for choice B will tend also to have a higher than predicted proportion of students opting for choice C. This is reflected in the residual rankings below. We found no school-level residual variance in the gender effect, so the residuals are associated with the intercept terms. The full tabulation of the random part of Model 11 is:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ^2_v(2_cons)	0.1462	0.0458	1
σ_v(3_cons,2_cons)	0.2202	0.05517	0.974
σ^2_v(3_cons)	0.3492	0.1106	1
σ_v(4_cons,2_cons)	0.1116	0.04565	0.591
σ_v(4_cons,3_cons)	0.1675	0.07182	0.574
σ^2_v(4_cons)	0.2441	0.08777	1
Extra-multinomial	0.9785	0.01621	

School-level residuals plotted against their ranks (Model 11):



School-level residuals, with their ranks (Model 11):

School	2_cons	3_cons	4_cons	2_rank	3_rank	4_rank
1	-0.166	-0.236	-0.332	23	23	15
2	0.325	0.475	-0.203	52	52	21
3	-0.262	-0.384	-0.534	17	17	3
4	-0.366	-0.584	-0.143	7	7	24
5	0.056	0.086	0.255	37	36	46
6	-0.348	-0.532	-0.324	8	8	16
7	-0.034	-0.025	-0.026	30	31	29
8	-0.054	-0.081	0.053	28	30	34
9	-0.025	-0.022	-0.461	33	32	8
10	-0.117	-0.186	-0.013	25	25	31

11	-0.320	-0.472	-0.138	9	12	25
12	-0.061	-0.106	-0.039	27	27	28
13	0.302	0.469	0.503	51	51	55
14	0.023	0.062	0.023	35	35	32
15	0.063	0.098	-0.280	38	38	19
16	0.280	0.406	0.522	50	50	56
17	-0.036	-0.083	0.153	29	29	40
18	-0.402	-0.623	-0.323	3	4	17
20	-0.303	-0.474	-0.400	13	11	13
21	0.135	0.162	0.753	44	41	60
22	-0.102	-0.151	-0.450	26	26	9
23	-0.308	-0.457	-0.530	11	14	4
24	0.219	0.306	0.312	49	47	49
25	-0.441	-0.646	-0.518	2	2	5
26	0.144	0.195	0.160	45	45	41
27	-0.190	-0.302	0.495	21	21	54
28	-0.285	-0.471	-0.344	15	13	14
29	0.567	0.866	0.572	57	57	58
30	0.461	0.716	0.324	54	55	50
31	0.127	0.182	0.122	43	43	37
32	-0.198	-0.334	0.242	19	19	44
33	0.112	0.169	-0.013	41	42	30
34	0.677	1.025	0.567	61	58	57
35	-0.452	-0.696	-0.612	1	1	1
36	0.356	0.579	0.246	53	53	45
37	-0.030	-0.010	-0.058	32	33	27
38	0.663	1.029	0.828	59	60	63
39	-0.368	-0.587	-0.448	6	6	10
40	-0.315	-0.478	-0.490	10	9	7
41	0.472	0.701	0.354	56	54	52
42	-0.248	-0.371	-0.312	18	18	18
43	0.667	1.028	0.754	60	59	61
44	-0.031	-0.099	0.134	31	28	39
45	-0.156	-0.187	-0.171	24	24	22
46	-0.395	-0.592	-0.552	5	5	2
47	-0.304	-0.476	-0.406	12	10	12
48	-0.398	-0.624	-0.494	4	3	6
49	0.016	0.001	-0.088	34	34	26
50	-0.196	-0.323	-0.159	20	20	23
51	0.714	1.099	0.415	62	62	53
52	0.655	1.046	0.289	58	61	48
53	0.097	0.126	0.264	40	40	47
54	1.048	1.617	0.679	63	63	59
55	-0.288	-0.449	-0.432	14	15	11
56	0.051	0.092	0.128	36	37	38
57	0.158	0.220	0.104	46	46	35
58	0.176	0.345	0.169	47	49	42
59	0.469	0.732	0.350	55	56	51
60	-0.268	-0.397	-0.238	16	16	20
61	-0.173	-0.287	0.029	22	22	33
62	0.123	0.193	0.195	42	44	43
63	0.217	0.309	0.805	48	48	62
64	0.072	0.121	0.110	39	39	36

Top-ranking schools on each category are (all significantly above expectation):
2_cons (likelihood of choosing option B): 54, 51, 34, 43, 38, 52, 29, 41, 59, 30, 36, 2
3_cons (likelihood of choosing option C): 54, 51, 52, 38, 43, 34, 29, 59, 30, 41, 36, 2
4_cons (likelihood of choosing option D): 38, 63, 43, 21.

We now present two models that attempt to describe what is happening at student level.

5.2 Models for explanation

Model 12

In this model we adapt a technique used by H&H, of estimating the effects on the probabilities of each choice of

- the student's belief that that choice would obtain best mark
- the student's view of the explanatory power of that choice
- the correctness of the student's validity rating of that choice.

Note that we do not estimate the effect on the probability of choice C of the student's validity rating of choice B, for example. And the student's opinions about best mark, explanatory power, and validity for choice A appear nowhere in the model.

The random parts of this model and of Model 13 are of no interest, and we do not discuss them. The fixed-part prediction is:

$$\begin{aligned}
\text{predicted } \ell_2 &= \quad -0.15base - 0.45girl + 0.25best + 0.41\exp l - 0.35valid, \\
\text{predicted } \ell_3 &= -2.27 - 0.15base - 0.82girl + 1.49best + 0.31\exp l + 0.46valid, \\
\text{predicted } \ell_4 &= -1.43 \quad\quad - 0.79girl + 1.84best + 0.65\exp l - 0.96valid,
\end{aligned}$$

with standard errors as below:

PARAMETER	ESTIMATE	S. ERROR
3_cons	-2.268	0.2007
4_cons	-1.433	0.1472
2_base	-0.148	0.04425
3_base	-0.1543	0.06758
2_girl	-0.4526	0.07974
3_girl	-0.8236	0.1342
4_girl	-0.7931	0.133
2_2best	0.2463	0.09311
3_3best	1.488	0.1742
4_4best	1.84	0.1333
2_2expl	0.4125	0.07122
3_3expl	0.3096	0.1465
4_4expl	0.6469	0.1342
2_2valid	-0.346	0.08141
3_3valid	0.4634	0.1303
4_4valid	-0.964	0.2258

Girls are less likely than boys with the same other characteristics to choose any of B, C, and D. The coefficients of *best* and of *expl* are all positive, indicating that a belief that one of these options either would gain best mark or has explanatory power (or both) predisposes the student to choose that option. A correct assessment of the validity of options B and D makes choosing those options less likely – which is rational, since they do not represent valid arguments – while the best choice, C, is made more likely by a correct validity assessment. A high baseline score is associated with *lower* likelihood of choosing the best option, conditionally on the other effects in this model. (Recall that, without the effects *best*, *expl*, and *valid*, baseline score has no significant effect on the likelihood of choosing option C.) Total score for constructive proof in geometry, and proof-as-general (L1b) score, are not statistically significant once the above effects are included.

Model 13

We tested the significance of constructive scores in Geometry when added to Model 12 and found them to be non-significant. If we attempt to include the overall validity rating (VR) score in Geometry as well as the student's validity rating of the particular choice made, we find, first, that the additional predictive power of total VR score is non-significant for choice B. Secondly, its inclusion makes the validity rating for choice C non-significant as a predictor for choice C. Thirdly, a high overall VR score is estimated to make choice D more likely, while at the same time a correct assessment of the validity of option D makes that choice very much *less* likely. This combination is difficult to interpret.

Accordingly, in Model 13, we include standardised overall VR score in Geometry as a predictor for choice C, but retain correct validity assessment as predictor for choice B and D, respectively. The distribution of Geometry and Algebra VR scores was given in Section 2.2. The fixed-part prediction of Model 13 is:

$$\begin{aligned}
\text{predicted } \ell_2 &= \quad -0.16base - 0.46girl + 0.24best + 0.42\exp l - 0.34valid \\
\text{predicted } \ell_3 &= -2.24 - 0.19base - 0.88girl + 1.47best + 0.34\exp l + 0.28VRscore + 0.38lotsofmaths, \\
\text{predicted } \ell_4 &= -1.43 \quad\quad\quad - 0.79girl + 1.84best + 0.65\exp l - 0.96valid,
\end{aligned}$$

with standard errors as below:

PARAMETER	ESTIMATE	S. ERROR
3_cons	-2.239	0.206
4_cons	-1.432	0.1479
2_base	-0.1554	0.04485
3_base	-0.1947	0.06998
2_girl	-0.4584	0.08048
3_girl	-0.8826	0.1345
4_girl	-0.7945	0.134
2_2best	0.2441	0.09392
3_3best	1.471	0.1746
4_4best	1.838	0.1342
2_2expl	0.4176	0.07194
3_3expl	0.3405	0.1441
4_4expl	0.649	0.1352
3_vrscore	0.2764	0.0655
3_lotsofmaths	0.3807	0.1914
2_2valid	-0.3388	0.08422
4_4valid	-0.9624	0.2292

The effects of *base*, *girl*, *best*, and *expl* are similar to Model 12. A high overall VR score is associated with higher probability of choosing option C. Correct assessments of the validity of options B and D make these choices less likely. There is an additional positive school-level effect on this choice of more than three hours of maths a week (*lotsofmaths*). This effect just fails to reach statistical significance at the 5% level in Model 12.

6 Choice for own approach in question A3

As in G3, there are four possible choices, coded 1, 2, 3, and 4 for A, B, C, and D. Choice A (1) is considered *empirical* and is used as the base category. The same set of variables was tested for significance as for G3, with algebra scores substituted for geometry as appropriate. The pattern was similar: the school-level variable found to be significant was 'lotsofmaths', i.e. more than 3 hours per week. Textbook 2 was initially found to have a significant negative effect on choice D, but ceased to be significant in the presence of other variables. No class-level variable was significant.

The fixed-part prediction is:
predicted $\ell_2 = -0.2752$,
predicted $\ell_3 = -0.2257$,
predicted $\ell_4 = -1.466$,
with standard errors as under:

PARAMETER	ESTIMATE	S. ERROR
2_cons	-0.2752	0.05443
3_cons	-0.2257	0.0425
4_cons	-1.466	0.08922

Thus, all other choices are significantly less likely than the empirical.

There is no statistically significant school-level residual variation in the probability of choice C. The residual variance/correlation matrix at this level is:

	B	D
B	0.066	
D	$r = 0.75$	0.186.

The full tabulation of the random part of the intercept model is:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(2_cons)	0.06632	0.02957	1
σ_v(4_cons,2_cons)	0.08309	0.03577	0.749
σ_v^2(4_cons)	0.1856	0.08508	1
Extra-multinomial	0.9884	0.01567	

As before, we present first a model for school comparison, and then elaborate it to explain at student level.

6.1 A model for comparing schools

Model 14

With the addition of gender and baseline score effects, school-level variation is reduced to a small variance (0.078) in the probability of choosing option B. Accordingly, we shall be able to produce residual rankings on this outcome only. The fixed-part prediction is:

$$\text{predicted } \ell_2 = -0.57\,girl,$$
$$\text{predicted } \ell_3 = -0.40\,girl - 0.14\,base,$$
$$\text{predicted } \ell_4 = -1.14 - 0.79\,girl - 0.53\,base,$$

with standard errors as under:

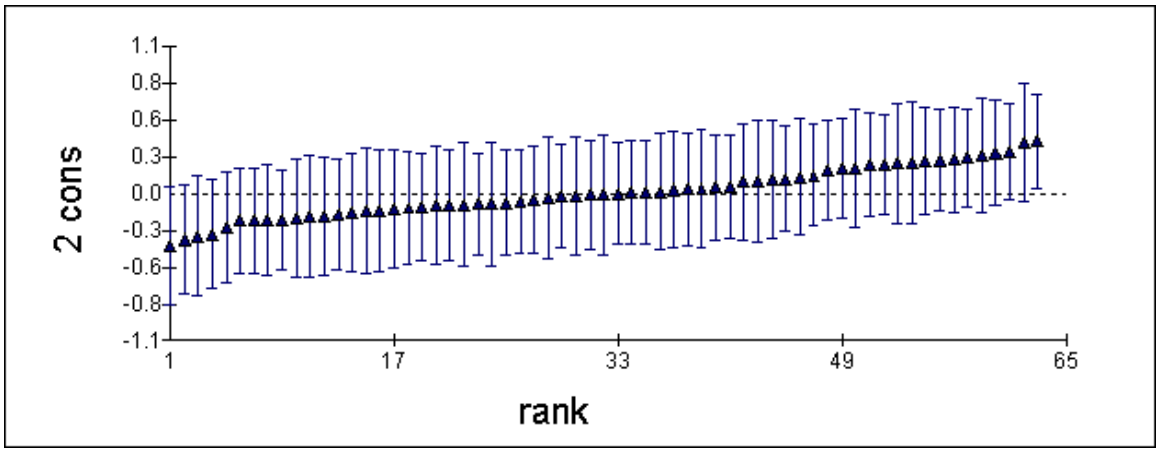
PARAMETER	ESTIMATE	S. ERROR
4_cons	-1.136	0.09099
2_girl	-0.5729	0.07296
3_girl	-0.398	0.06267
4_girl	-0.7921	0.1418
3_base	-0.1381	0.04184
4_base	-0.5348	0.06729

There is no baseline effect on option B (the best option). A high baseline score is associated with a lower probability of choosing C or D. Boys at about the mean of baseline score are as likely to choose B or C as they are to choose A, and less likely to choose D. As with G3, we find that girls are more likely than boys with similar other characteristics to choose the 'empirical' option, A.

The random part of Model 14 is:

PARAMETER	ESTIMATE	S. ERROR	CORR.
School			
σ_v^2(2_cons)	0.07834	0.032	1
Extra-multinomial	1.001	0.01636	1.001

This gives the following plot of school-level residuals for choice B:



School-level residuals for choice B plotted against their ranks

Only one school, number 62, is significantly above expectation in the proportion of its students choosing option B. No school is significantly below expectation. The full residual listing is on the following page:

School-level residual ranks for option B (Model 14):

Rank	Resid	School
63	0.389	62
62	0.380	31
61	0.310	2
60	0.304	38
59	0.286	13
58	0.264	43
57	0.253	46
56	0.250	50
55	0.238	54
54	0.227	53
53	0.226	32
52	0.215	59
51	0.213	29
50	0.186	40
49	0.186	14
48	0.177	12
47	0.135	4
46	0.124	16
45	0.111	52
44	0.103	17
43	0.088	30

Rank	Resid	School
42	0.085	63
41	0.054	8
40	0.044	49
39	0.042	55
38	0.031	36
37	0.029	34
36	0.013	33
35	0.007	7
34	0.004	35
33	-0.003	6
32	-0.007	1
31	-0.010	3
30	-0.020	48
29	-0.024	27
28	-0.031	56
27	-0.047	39
26	-0.060	58
25	-0.068	20
24	-0.079	15
23	-0.079	45
22	-0.082	37

Rank	Resid	School
21	-0.086	9
20	-0.091	18
19	-0.102	23
18	-0.108	11
17	-0.113	26
16	-0.128	51
15	-0.130	5
14	-0.142	10
13	-0.155	57
12	-0.167	41
11	-0.169	64
10	-0.182	22
9	-0.193	44
8	-0.198	47
7	-0.200	60
6	-0.201	21
5	-0.252	25
4	-0.303	28
3	-0.316	61
2	-0.343	24
1	-0.389	42

6.2 A model for explanation

Model 15

This model, for students' own approaches to A3, was developed along similar lines to Model 13 for own approaches G3. The fixed-part prediction is:

$$\text{predicted } \ell_2 = -0.60 - 0.63girl - 0.14base + 0.59best + 0.30valid + 0.40lotsofmaths + 0.31VRscore,$$
$$\text{predicted } \ell_3 = -0.69 - 0.41girl - 0.14base + 1.55best + 0.74\exp l - 1.12valid,$$
$$\text{predicted } \ell_4 = -1.20 - 0.72girl - 0.46base + 1.52best + 0.56\exp l - 0.64valid,$$

with standard errors as in the following table:

PARAMETER	ESTIMATE	S. ERROR
2_cons	-0.6019	0.1075
3_cons	-0.6888	0.09681
4_cons	-1.204	0.1272
2_girl	-0.6293	0.09754
3_girl	-0.4056	0.09838
4_girl	-0.7232	0.1516
2_base	-0.1439	0.05229
3_base	-0.1358	0.05055
4_base	-0.4564	0.07295
2_2best	0.5925	0.09958
3_3best	1.55	0.09677
4_4best	1.515	0.2125
3_3expl	0.7403	0.09262
4_4expl	0.5617	0.145
2_2valid	0.2986	0.1039
3_3valid	-1.118	0.168
4_4valid	-0.6376	0.1676
2_lotsofmaths	0.3951	0.1153
2_vrscore	0.3141	0.05465

The most interesting aspect of this model is the prediction for option B. Explanatory power is not a statistically significant predictor when correct assessment of validity is already in the model, but a correct assessment of the validity of this proof, more than 3 hours of maths a week, and a high overall VR score in algebra all play a significant part in disposing the student to make this choice. We found that adding the effect of Algebra constructive scores to this model yielded results that were difficult to interpret. A high score for constructive proof was apparently associated with *reduced* probability of making the best choice (B) for own approach to A3; its effect on choice C was non-significant, and on choice D negative. Accordingly, we exclude from the Model 15 the effect of Algebra constructive scores.

In the random part of this model, which we do not show, no significant residual school-level variation remains.

7 Choice for best mark in question G3

7.1 A model for explanation

As with choice for own approach in G3, a correct validity rating for a particular option has an apparently rational effect on the probability of choosing it for best mark. We find that a high overall VR score in Geometry has a further significant positive effect on the likelihood of choosing proof C (the best option). Attempting to model all choices on VR score without individual validity ratings leads to instability: there is no clear connection between VR score and choices B and D. The predicted effect of constructive scores in Geometry is uninteresting: a high constructive score is predicted to increase the probabilities of choosing each of B, C, and D for best mark (by comparison with A), but the smallest effect, which only just reaches statistical significance, is on the probability of choice C. We omit these effects from Model 16.

Model 16

The fixed-part prediction is:

$$\begin{aligned}
\text{predicted } \ell_2 &= 0.55 + 0.69\exp l - 0.59 valid,\\
\text{predicted } \ell_3 &= 1.09 - 0.33 girl + 0.94\exp l + 0.27 valid + 0.22 VRscore + 0.47 text2,\\
\text{predicted } \ell_4 &= -0.40 girl + 0.81\exp l - 1.36 valid,
\end{aligned}$$

with standard errors as in the following table:

PARAMETER	ESTIMATE	S. ERROR
2_cons	0.5486	0.0997
3_cons	1.091	0.09882
2_2valid	-0.5875	0.09533
3_3valid	0.2714	0.06907
4_4valid	-1.363	0.2351
3_girl	-0.3335	0.06734
4_girl	-0.3994	0.1129
2_2expl	0.6944	0.07496
3_3expl	0.9419	0.06268
4_4expl	0.8064	0.09017
3_text2	0.4654	0.221
3_vrscor	0.2199	0.03792

Note that baseline score does not significantly affect choice for best mark in G3. A correct validity rating disposes the student in favour of choosing option C for best mark and against options B and D, as would be expected. Girls are less disposed than boys with similar other characteristics to choose options C and D, but there is a net preference for C as opposed to A, even among those who do not correctly judge its validity and score poorly overall on VR. Interestingly, textbook 2 tends to dispose students to choose option C.

8 Choice for best mark in question A3

8.1 A model for explanation

As with choice for own approach to A3 (Model 15), we found that correct validity ratings for individual options had rational effects on the probabilities of choosing those options for best mark. In addition, a high overall VR score for Algebra increased the probability of choice B (the best choice). With individual validity ratings already in the model, VR score made no additional contribution to predicting the probability of choice D, and in the case of choice C there was a contradiction: a high overall VR score seemed to favour this choice, but a correct validity rating (as expected) made the choice less likely. Accordingly, we exclude VR score from the model for choices C and D.

Model 17

The fixed-part prediction is:

$$\begin{aligned}
\text{predicted } \ell_2 &= 0.58 + 0.47valid + 0.42\exp l + 0.22VRscore,\\
\text{predicted } \ell_3 &= -0.25girl - 1.29valid + 0.89\exp l,\\
\text{predicted } \ell_4 &= -1.28 - 0.46girl - 0.99valid,
\end{aligned}$$

with standard errors as below:

PARAMETER	ESTIMATE	S. ERROR
2_cons	0.583	0.07668
4_cons	-1.284	0.172
3_girl	-0.2514	0.07057
4_girl	-0.4621	0.2168
2_2valid	0.4728	0.076
3_3valid	-1.294	0.1613
4_4valid	-0.9897	0.2606
2_2expl	0.4245	0.07115
3_3expl	0.8907	0.06606
2_vrscor	0.2216	0.03978

A perverse effect by which a high baseline score was predicted to favour choice C has been excluded. There were no other statistically significant effects of baseline score on the choice for best mark in A3. Girls are as likely as boys with similar other characteristics to choose option B (the best choice), but less likely than boys to choose C or D. Option D is the least likely choice of either boys or girls, even if its validity is incorrectly assessed. The effect of constructive score in Algebra is uninteresting: a high score is predicted to make option D less likely, but the effect is otherwise non-significant. Accordingly, we omit this effect from Model 17.

9 Exploring the school-gender effect

When the school-gender codes are corrected, a school-gender fixed effect on constructive proof in Geometry is detected in Models 1, 3, and 6, but not in Models 2, 4, 5, or 7. The latter models include a fixed effect of school % A*–C at GCSE; Models 1, 3, and 6 omit this effect.

For illustration, fixed-part estimates are tabulated below for a model including all the effects of Model 6, together with fixed effects of school gender on the four outcomes, Geometry (prefixed `geo_`), Algebra (`alg_`), Geometry VR (`geoVR_`), and Algebra VR (`algVR_`)

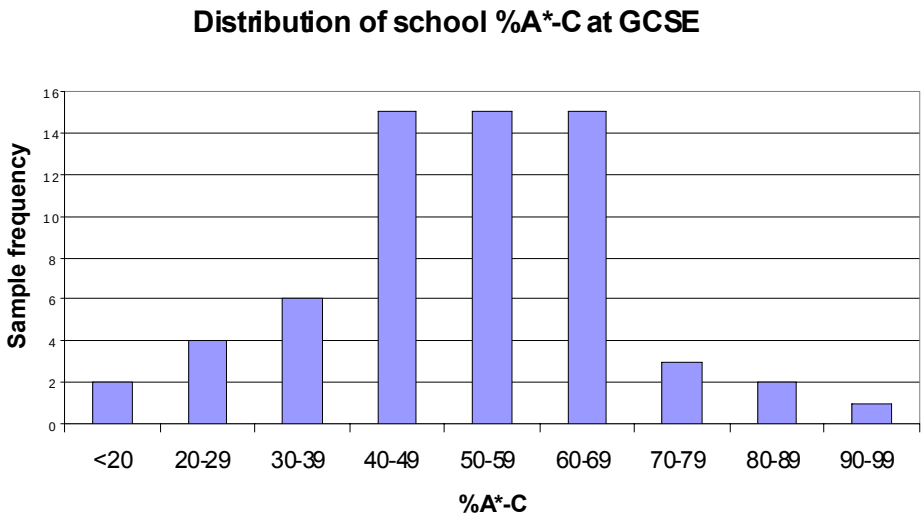
PARAMETER	ESTIMATE	S. ERROR
geo_cons	7.525	0.1494
alg_cons	8.877	0.1988
geoVR_cons	2.473	0.05918
algVR_cons	2.755	0.07259
geo_girl	0.3782	0.1241
alg_girl	0.5817	0.168
geoVR_girl	0.05457	0.0705
algVR_girl	0.149	0.07839
geo_base	1.532	0.09612
alg_base	1.874	0.1209
geoVR_base	0.5108	0.05598
algVR_base	0.466	0.05855
geo_base2	0.2219	0.06078
alg_base2	0.4766	0.07644
geoVR_base2	0.1825	0.03642
algVR_base2	0.1312	0.0376
geo_base3	0.05961	0.0313
alg_base3	0.1577	0.03936
geoVR_base3	0.03973	0.01862
algVR_base3	0.04558	0.01929
geo_girls_school	**1.353**	**0.5201**
alg_girls_school	**0.3986**	**0.5486**
geoVR_girls_school	**0.2483**	**0.1696**
algVR_girls_school	**0.3791**	**0.2344**

(Compare with the tabulation on p29 of the report.)

Note that this model includes the outcomes of Models 1 and 3. It can be seen from the standard errors of the estimates that the school-gender effect (suffixed `girls_`school) is statistically significant only for Geometry constructive proof and not for Algebra constructive proof, Geometry VR, or Algebra VR. The effect disappears when school % A*–C at GCSE is introduced into the model.

The question is whether this result should be regarded as indicating a beneficial effect of girls' schools on Geometry constructive proof which is general in its applicability.

The four girls' schools in the sample, numbers 1, 5, 30, and 31, came respectively 16th, 24th, 1st, and 7th, based on their residuals for girls' Geometry scores from Model 1. Thus, all four schools were ranked in the top half of the sample by that model, with two performing significantly above the mean. When GCSE % A*–C was introduced, in Model 2, their positions became respectively 34th, 18th, 20th, and 7th, and none of them performed significantly above the sample mean.



The diagram shows the sample distribution of school GCSE %A*-C.

The figures for the girls' schools are:

School	1	5	30	31
GCSE %A*–C	84	38	99	61

Schools 30 and 1 are respectively 1st and 2nd in the sample on historical GCSE performance. Either they are outstanding in their preparation of students for GCSE, or they have a very favoured intake, or both. It is these two schools that are responsible for the apparent effect of girls' schools on Geometry scores in models that exclude the effect of GCSE score. It is not surprising that they drop substantially in the residual rankings when GCSE *A–C is included as a fixed effect, and that the apparent effect of girls' schools on Geometry proof then disappears. With so few girls' schools in the sample, and two of these probably untypical, there is no evidence for a general school-gender effect on Geometry proof.

For the same reason, any other effects of school gender, for example interaction effects or effects on student choices in A3 and G3, that may be statistically detectable for this sample will not be generalisable.

10 Summary

10.1 Total scores

In sections 2 and 3 we modelled total scores for Geometry and Algebra constructive proof, and for Geometry and Algebra validity rating (G3c and A3c, respectively), as multivariate responses. The scores for the 'Logic' question , L1, were included in the total for Algebra constructive proof.

The student's score on the baseline test was a statistically significant predictor for all four scores.

For constructive proof, we found (Model 0) that without adjustment for baseline score there was no statistically significant effect of gender. When baseline score was included in the model (as it was in all other models), there was in addition a statistically significant gender effect, in favour of girls, except in the case of Geometry VR score (see Model 6). This gender effect, where present, was not itself dependent on the student's baseline score, in other words, there was no interaction between gender and baseline score. Girls tended to perform less well than boys on the baseline test: for example, the median baseline score for girls is 15 and for boys is 16. The gender effect, where present, was comparable in size and of opposite sign to the effect of this one-point difference in baseline score. In Geometry VR score, where no statistically significant gender effect was found, the effect of a one-point difference in baseline score was at most one-tenth of a raw-score point. See the table on page 45 for more detail.

We may summarise this finding by saying that girls tended to perform better on these proof tests, relatively to their performance on the baseline test, than boys.

Also in Model 0, we found slight but statistically significant variation in constructive proof scores between classes within schools, amounting to some 3 to 4 per cent of the total variance (between-school variance accounted for five times as much as this). Once baseline score was adjusted for, variation between classes within schools ceased to be statistically significant. Not could we find any teacher-level variables with statistically significant effects.

The only variables not already part of the proof test that were found to have a statistically significant effect on total scores, for either constructive proof or validity rating, were:

School level
- %A*-C
- Use of Textbook 2

Class level
- (none)

Student level
- Gender
- Baseline score

In particular, whether a school was 11–18 or girls-only had no statistically significant effect. Textbook 2 was found to be beneficial for Algebra constructive scores and for Geometry VR scores. The school's %A*-C had a positive effect on all four scores.

After adjustment for student gender, baseline score, and the school's %A*-C, we found that schools varied in their effectiveness, except in the case of Geometry VR score, for which no school-level variation was detected. Schools' residual performance in Algebra constructive proof was significantly more variable for boys than for girls once the school's %A*-C was adjusted for. Correlations at school level between girls' and boys' residual performance were high ($r \cup 0.9$)

within Algebra and Geometry constructive proof and appreciable ($r > 0.7$) for Algebra VR, but correlations across these subjects were more modest (generally, $r < 0.6$). Thus, in a school whose girls did better than predicted in Algebra the boys tended to do better in Algebra also; but neither the girls nor the boys in that school would be especially likely, as a group, to do better than predicted in Geometry. This is demonstrated by the rankings of schools for Algebra and Geometry in Models 1 and 2.

We found no evidence of differences in school effectiveness for students with different baseline scores.

Turning to the relative sizes of the different effects, we need to be careful not to over-interpret apparent differences. All the estimates are subject to error. With this caveat, we found from Model 7 that the gender effect on total Algebra constructive score was about half the effect of using textbook 2. The effect on this score of the school's %A*-C, compared to the average, was less than the gender effect for more than 70% of students. The remaining 30% of students attended schools whose %A*-C departed so far from the mean that its effect equalled or exceeded the gender effect. By contrast, the effect of baseline score for the top 15% and the bottom 15% of students was of a higher order of magnitude than the gender effect. Residual school effects on the total Algebra constructive score for girls were generally smaller in size than the gender effect itself, though schools at the extremes exceeded this. For boys, as we have noted, the school effects tended to be larger. See the tables on pages 43 and 45 for more detail and for the effect sizes for the other scores.

10.2 Individual scores for constructive proof

These were modelled in section 4, and no attempt was made to rank schools on these very limited outcomes. The main purpose was to study correlations between scores on different questions and whether any additional statistically significant effects could be found. We consider the Geometry questions first.

Student baseline score and school %A*-C were significant predictors for all individual Geometry scores. Gender was significant for all except G2a, but *negative* for G1. Geometry VR score was significant for all except G2b. The response coded 30 to question L1b was interpreted as an indicator that the student appreciated proof as general. This indicator had a statistically significant positive effect on the score for G4 (only). Use of textbook 2 was found to have a statistically significantly positive effect on G1 score (only).

We could find no statistically significant school-level residual variation in the scores for G2a or G2b. There was high residual correlation at school level between girls' and boys' scores for G1 and between girls' and boys' scores for G4; also between boys' scores for G1 and G4, but not between girls' scores for the separate questions. Correlations at student level were all low. An individual student who scores highly on one of these questions, compared to expectation and after adjusting for any school residual effect, is not especially likely to score highly on another.

We now consider the separate Algebra questions A1, A2, and A4, together with the Logic question, L1. Student baseline score and school %A*-C were significant predictors for all scores. Gender was significant, and positive, for A2 and A4 only. Algebra VR score was significant for A4 and L1 only. 'Proof-as-general' was significant for A2 and A4. Use of textbook 2 was significant for A1 and A2 only.

With these fixed effects in the model, statistically significant residual correlation at school level between girls' and boys' scores was present for question A4 only. Within gender, there appeared to be high school-level correlation between A1 and A2 scores, and between A2 and L1 scores. At student level, correlations were once again low.

Where these findings differ from those for the total constructive scores they are not easy to interpret. And all should be treated with caution in view of the distributional characteristics of the individual scores. See pages 48, 53, and 54.

10.3 Choice of own approach in multiple choice questions

Probabilities of different choices for the student's own approach in these two questions, as indicated by their answers to G3a and A3a, were analysed in sections 5 and 6. No class-level residual variation was found, and no statistically significant class-level or teacher effects. In particular, the teacher's choice for their own approach was not significant.

The following variables, not already part of the proof test, were found to have a statistically significant effect on at least one of these probabilities:

School level

More than 3 hours of maths per week

Student level

Gender

Baseline score

Note that neither school %A*-C nor use of a particular textbook nor GCSE syllabus was significant. While the amount of maths taught per week was significant, the existence of a maths club was not.

Models of two types were considered. The first type was for school comparisons, and excluded all variables apart from those tabulated above. The second type was designed to explore how the outcome probabilities were associated with other aspects of the student's response to the proof test.

In a model to compare school effects, the number of hours of maths per week should not be included: although significant, it is, like choice of textbook, within the school's control. This leaves only gender and baseline score.

We consider G3 first. Of the four possible choices, A was termed 'empirical', and was used as the base category with which to compare the probabilities of each of the other three choices. Choice C was a correct formal proof and was objectively the best approach. Interestingly, baseline score had *no* statistically significant effect on the probabilities of choices C or D: both were significantly less likely than choice A. Baseline score had a significant negative effect on the probability of choice B: thus, this choice became less likely, the higher the baseline score. Given a baseline score, girls were less likely than boys to make choice B, and they were always less likely than boys to choose C or D.

We found no significant school-level residual variance in the gender effect. There was very high correlation ($r > 0.9$) at school level between choices B and C, in other words, a school with a higher than predicted proportion of students opting for choice B would tend also to have a higher than predicted proportion opting for choice C. Correlations for the other pairs of choices (B,D and C,D) were only moderate ($r < 0.6$).

Following Healy and Hoyles (1999), we then explored the effects on the probabilities of choosing B, C, and D of the student's view of that proof's explanatory power and validity, and whether it would gain the best mark from the teacher. We also allowed total Geometry VR score and 'more than 3 hours of maths per week' as possible predictors. We found no significant effect of scores in Geometry constructive proof. For each choice, B, C, or D, we found that thinking a proof would gain the best mark predisposed a student to choose it for their own approach, as did thinking it explained *why* the result was true (or false). A correct assessment of the proof's validity (as shown by the student's response to question G3c) made choices B and D significantly less likely. (These proofs are not valid.) A correct assessment of validity for choice C (the best proof) made it more

likely, but more significant for the probability of this choice was the total VR score in Geometry (i.e. the total score for G3c). Being in a school that offered more than three hours of maths a week also had a statistically significant positive effect on the probability of choosing option C. This effect of numbers of hours of teaching, which was confined to the probability of option C, was present in other models not described in this report, and was the only school-level fixed effect found to be statistically significant for this outcome.

We now consider question A3 – the Algebra multiple choice question. Again, there were four possible choices, A to D, and we termed choice A 'empirical' and used it as the base. The best choice (in the sense of most complete and valid) was option B. Baseline score had no statistically significant effect on the likelihood of choosing this proof, which boys were as likely to choose as option A. Girls, whatever their baseline score, were significantly less likely than boys to choose option B. The likelihood of choices C and D diminished with increasing baseline score and was less for girls than for boys with a given baseline score.

After adjustment for gender and baseline score, there was little residual variation at school level, and this only in the probability of choosing option B. Only one school – number 62 – was significantly above expectation in the proportion of its students choosing this option.

A correct assessment of the validity of option B increased the probability of the student's choosing it for their own approach. Correct assessments of validity for options C and D (which are invalid) reduced the probability of their choice. Thinking that a proof, whether B, C, or D, would gain the best mark from the teacher increased the likelihood of its choice as the student's own approach. The likelihood of choosing option B appeared to be unaffected by the student's opinion of that proof's explanatory power, while a positive view of the explanatory power of either proof C or proof D increased the likelihood that the student would choose it for their own approach. As was the case with the objectively best choice for Geometry, we found that more than three hours of maths per week in the school had a statistically significant and positive effect on choice B for Algebra. A high total VR score in Algebra also was associated with increased likelihood of this choice as the student's own approach. In the presence of these rational effects, the effects of Algebra constructive scores proved to be difficult to interpret. These were, therefore, omitted.

10.4 Choice for best mark in multiple choice questions

Probabilities of different choices for best mark for these two questions, as indicated by students' answers to G3b and A3b, were analysed in sections 7 and 8. As in the previous analyses, choice A was in each case the base. As before, no class-level residual variation was found, and no statistically significant teacher effects. In particular, the teacher's choice for their own approach was not significant for student choices for best mark. Nor was teacher's opinion of what students would choose for best mark.

The variables that were not already part of the proof test and were found to have a statistically significant effect on at least one of the probabilities in this group were:

School level
Use of Textbook 2

Student level
Gender.

Thus, there was no effect of baseline score.

We were not concerned to compare schools on these outcome probabilities, so we proceeded directly to explore associations with other student responses to the proof test, as well as the variables above. We excluded student choice for own approach as a predictor, as the previous

analyses had already demonstrated that the two probabilities were related, and it is more reasonable to suppose that the student's view of what would gain the best mark should predict the student's choice for their own approach, rather than the other way round.

Considering Geometry first, we found that the probability of choosing proof B, C, or D in preference to A for best mark was increased if the student considered that proof to explain the result. A correct assessment of the validity of the proof increased the likelihood of its choice for proof C (the best proof) and decreased it for proofs B and D. A high total VR score for Geometry further increased the probability of choosing proof C for best mark. Interestingly, use of textbook 2 also increased the probability of choosing proof C. In fact, there was a preference for option C over option A for best mark even among those who did not correctly judge its validity and had poor overall VR scores. Girls were less likely than boys to choose proofs C or D, conditionally on the other variables.

All of these effects (apart from the gender effect) are relatively easy to understand. By contrast, the effect of constructive proof scores in Geometry was not easy to interpret. High constructive scores were predicted to make each of choices B, C, and D more likely for best mark. The effect on choice C, however, only just reached statistical significance and had the smallest magnitude. It seems most reasonable to ignore these apparent effects.

In Algebra, proof D was by far the least likely to be chosen for best mark, and its likelihood was further decreased by a correct assessment of its validity. Girls were less likely than boys to make this choice, as they were also to make choice C. Proof B, objectively the best, was generally most likely to be chosen for best mark, and this probability was further increased if the student correctly judged its validity or felt it had explanatory power. A high total VR score in Algebra further increased the probability of this choice for best mark. Proof C also was made more likely if the student felt it had explanatory power, but less likely if the student correctly judged its validity. The effect of constructive score in Algebra was not of interest: a high score was predicted to make option D even less likely, but the effect was otherwise non-significant.

References

Goldstein, H. (1995) *Multilevel Statistical Models*, 2Ed (London: Edward Arnold)

Healy, L. and Hoyles, C. (1999) *Justifying and Proving in School Mathematics: Technical report on the nationwide survey* (London: Institute of Education, June)