Chapter 7
Mathematics achievement

This chapter provides information on pupil mathematics achievement at the start of the evaluation in September 2013 and in May-June 2014.[1] The main function of the chapter is to examine the extent and nature of change in mathematics achievement (if any) between September and May within each programme, overall, and for subsets of pupils.

The chapter is divided into seven sections, the first of which deals with overall mean scores of pupils, by programme. Section two analyses results by content and process. Section three examines achievement results by gender, while the fourth section examines how high- and low-achieving pupils performed in each programme. Section five examines pupil achievement by assigned observation type (i.e., recorded or live observations). The sixth section examines how teacher and teaching characteristics (i.e., teachers' mathematical knowledge for teaching, mathematical quality of instruction, and programme adherence) relate to pupil achievement, and if teachers' mathematical knowledge for teaching changed during the evaluation. Finally, the relationship between pupil achievement and attitudes to mathematics is examined. The analyses reported in the chapter are unadjusted for multiple comparisons, but where adjusted *p* values alter significance, this is noted at the end of the chapter.

Overall score on the DPMT

In September 2013, 546 pupils completed the DPMT, Level 2 (autumn norms). In May 2014, 536 pupils completed the DPMT, Level 3 (spring norms). In total, a "core group" of 509 pupils (271 in JUMP and 238 in IMPACT) completed the test on both occasions. Unless otherwise stated, all analyses that follow refer to the core group, or to a relevant subgroup within it. This allows a like-with-like comparison, as the same group of pupils are being compared on each occasion. However, readers should note that the mean score for pupils in the core group was slightly higher (on both occasions) than the mean score for all pupils sitting the test. This is an expected finding, as pupils absent for testing generally tend to have slightly lower achievement than those present for testing (see Cosgrove, 2005, for a discussion of the issue).

In September 2013, pupils in the core JUMP and IMPACT groups achieved mean standardised scale scores of 102, equivalent to the 55th percentile on the DPMT (Table 7.1). The similarity between the two groups was expected, given that standardised test scores from the end of the 2012/13 year had been used to assign schools to the two programmes in a way that maximised similarity in pupil achievement. Mean scores from the school-administered tests at Second class had averaged 109 in both groups, suggesting a drop in scores over the summer. However, this may be an artefact finding, due to a combination of different tests being used at the end of the previous year, pupil nervousness in September due to being tested by an external test administrator, the absence of a "halo effect"[2] when test questions were scored by ERC staff, and September testing taking place unusually early in the school year. This early timing

[1] As almost all pupils tested on the second occasion were tested in May, with only a small number tested during the first days of June, the second test period will hereafter be referred to as the May testing period.

[2] The halo effect refers (in this instance) to a tendency for teachers to give their pupils the benefit of the doubt on test items with ambiguous responses. Teachers may draw on what they already know about pupil knowledge of a topic and score accordingly, rather than scoring purely based on the response provided in the test.

avoided programme effects, but increased summer learning loss effects, which can be pronounced for mathematics (Cooper, Nye, Charlton, Lindsay & Greathouse, 1996). Irrespective of the cause of the change from May to September 2013, the most important aspect – that the two groups were well matched on average achievement – remained unchanged.

Table 7.1: Mean achievement scale scores (and standard deviations) for the core group of pupils, by programme and test time

	JUMP (N=271)	IMPACT (N=238)
Sept. 2013	102.0 (13.6)	102.2 (15.4)
May 2014	108.9 (14.4)	107.5 (17.8)
Change	+6.9 (8.8)	+5.3 (9.8)

The September 2013 mean scores of pupils in both programmes were slightly, but significantly, higher than the DPMT mean score of 100 (based on a national standardisation sample from 2006) ([JUMP: *t*(270)=2.36, *p*=.019]; [IMPACT: *t*(237)=2.15, *p*=.033]). The slight difference could be attributed to the self-selected nature of schools involved, to the fact that pupils absent on either test occasion were excluded from the mean, to improved mathematics achievement nationally since 2006, or (probably) to a combination of these factors.

In May 2014, mean scores for the core group of pupils increased in both programmes, to 109 for JUMP pupils (equivalent to the 73rd percentile), and to 107 for IMPACT pupils (equivalent to the 68th percentile). For both groups, the end-of-year results were again significantly higher than the DPMT mean of 100 obtained by the standardisation sample in 2005 ([JUMP: *t*(270)=10.13, *p*<.001]; [IMPACT: *t*(237)=6.46, *p*<.001)]).

Within each group, increases in pupil achievement over the course of the evaluation were significant ([JUMP: *t*(270)=-12.94, *p*<.001], [IMPACT: *t*(237)=-8.33, p<.001]). The standard deviation for the DPMT is 15 scale score points, meaning that the JUMP group improved by just under half a standard deviation, and the IMPACT group by roughly one third of a standard deviation. However, the difference in achievement scores *between* the JUMP and IMPACT groups was not significant at either test time ([September: *t*(507) =-0.15, *p*=.880], [May: *t*(507)=0.98, *p*=.328]). To correct for the fact that the samples were clustered by school, a Minimum Detectable Effect Size (MDES) was calculated (Hutchinson and Styles, 2010; Bloom, 1995) – i.e., the smallest gap that would produce a statistically significant difference between JUMP and IMPACT in the end-of-year tests.[3] The MDES was calculated at 6.1 scale score points. Since the JUMP and IMPACT groups differed by only 1.4 points in May, the difference between them was not statistically significant.

In sum, the achievement scores of JUMP pupils improved by an average of almost seven scale score points between the start and end of the evaluation, while those of IMPACT pupils improved by an average of just over five points. Pupils in each programme scored significantly higher than the DPMT mean of 100 on both occasions, and within each programme, the increase over the course of the evaluation was significant. However, pupils in the two programmes were not significantly different *from each other* in September (as was intended) or in May (when differences might have been expected if one programme was markedly more effective than the other).

[3] This calculation of the MDES assumes a two-tailed test, *p*<.05, power=80% (Bloom, 1995).

Performance on strands and content areas

The DPMT items are subdivided into four content groups that reflect the strands of the PSMC: Number and Algebra, Shape and Space, Measures, and Data. All items are also classified as requiring one of five processes: Understanding and Recalling, Implementing, Reasoning, Integrating and Connecting, and Applying and Problem-Solving. Standardised scale scores are not available at the content and process level, meaning we can only report *percent correct scores*. In practical terms, this means that content and process results should not be compared with each other or across Levels of the DPMT. For example, answering 66% of Data items and 55% of Measures items correctly at Level 2 is not equivalent to doing the same at Level 3, nor does it imply the pupil is better on Data than Measures. The Measures questions answered by the pupils may have been more difficult than the Data questions – something taken into account when reporting standardised scale scores, but not by a percent correct score. Therefore, content and process results are compared between JUMP and IMPACT groups in September and in May, but are not compared between the two times.

Content

At the start of the evaluation, the JUMP and IMPACT groups were very closely matched on the percent of items answered correctly, by content area (Table 7.2). The difference was largest for Shape and Space (an advantage of 2.4% in favour of IMPACT pupils) and smallest for Data (0.5% in favour of IMPACT).

Table 7.2: Pupils' mean percent correct scores by content strand and programme, **September 2013**

	JUMP	IMPACT	Gap
Number and Algebra	64.7	63.1	1.6%
Shape and Space	67.1	69.5	2.4%
Measures	45.5	47.7	2.2%
Data	50.1	50.6	0.5%

At the end of the year, the mean percentage of items answered correctly for three of four content areas was again very similar across the two programmes. The exception was Data, where JUMP pupils achieved a mean percent correct score 4% higher than that obtained by IMPACT pupils (Table 7.3). The difference on Data is nonetheless relatively small, and is almost entirely attributable to differences in performance on three items assessing pupil ability to interpret a pictogram. On these items, the percentage of JUMP pupils answering correctly was considerably higher than the percentage of IMPACT pupils who did so (differences of 22%, 13%, and 19%). Of course, as the Data strand was not covered in the available IMPACT manuals, it was probably taught to most IMPACT pupils using commercially available Irish textbooks.

Table 7.3: Pupils' mean percent correct scores by content strand and programme, **May 2014**

	JUMP	IMPACT	Gap
Number and Algebra	62.0	59.8	2.3%
Shape and Space	67.6	66.9	0.7%
Measures	56.9	55.2	1.7%
Data	66.9	62.9	4.0%

Process

At the outset, pupils' percent correct scores on all processes were very similar across the JUMP and IMPACT groups (Table 7.4). The largest difference was on Applying and Problem-Solving (a 2% gap in favour of IMPACT), while differences on Understanding and
Recalling, Reasoning, and Integrating and Connecting were all below 1%.

Table 7.4: Pupils' mean percent correct scores by process and programme, September 2013

	JUMP	IMPACT	Gap
Understanding and Recalling	66.3	66.9	0.6
Implementing	66.4	64.6	1.7
Reasoning	62.6	61.9	0.7
Integrating and Connecting	58.7	59.5	0.8
Applying and Problem Solving	45.3	47.3	2.0

At the end of the year, the groups were still closely matched across four of the processes. However, on Integrating and Connecting, JUMP pupils achieved a mean percent correct score almost 6% higher than that of IMPACT pupils (Table 7.5). Again, this can be partially attributed to large differences between JUMP and IMPACT pupils on the three items relating to pictograms (Integrating and Connecting was required for the three questions).

Table 7.5: Pupils' mean percent correct scores by process and programme, May 2014

	JUMP	IMPACT	Gap
Understanding and Recalling	72.8	71.2	1.6
Implementing	66.5	64.1	2.4
Reasoning	64.6	63.8	0.8
Integrating and Connecting	58.0	52.2	5.8
Applying and Problem Solving	53.5	51.5	2.0

Achievement by gender

As with the overall sample of pupils taking part in the evaluation, there were slightly fewer girls than boys in the core group tested (222 girls and 287 boys). Girls were outnumbered by boys in both JUMP (113 girls and 158 boys) and IMPACT (109 girls and 129 boys). Thus, there were fairly similar gender representations in the two groups, as girls composed 42% of pupils in JUMP and 46% in IMPACT.

In September, girls in both programmes achieved mean scores of 101, while boys in both programmes achieved mean scores of 103, a non-significant gender gap ($t(507)=-1.59$, $p=.112$). In May, boys again obtained slightly, but not significantly, higher mean scores than girls ($t(507)=-0.99$, $p=.322$). Girls and boys in JUMP achieved mean scores of 108 and 109, respectively, while girls and boys in IMPACT achieved scores of 107 and 108, respectively (Table 7.6). Thus, girls and boys in JUMP had an average increase of about seven points on the DPMT, while girls and boys in IMPACT both improved by between five and six points. These increases were significant for girls and boys in both programmes ([JUMP girls: $t(112)=-8.61$, $p<.001$], [JUMP boys : $t(157)=-9.64$, $p<.001$], [IMPACT girls: $t(108)=-6.15$, $p<.001$], [IMPACT boys: $t(128)=-5.67$, $p<.001$]).

Within each programme, gender differences were not significant in September ([JUMP: *t*(269)=-1.12, *p*=.261], [IMPACT: *t*(236)=-1.13, *p*=.259]) or May ([JUMP: *t*(269)=-0.61, *p*=.542], [IMPACT: *t*(236)=-0.73, *p*=.468]). Gender differences *across* programmes were also non-significant in September ([girls: *t*(220)=-0.03, *p*=.973], [boys: *t*(285)=-0.26, *p*=.792]) and May ([girls: *t*(208)=0.76, *p*=.449], [boys: *t*(245)=0.58, *p*=.563]). In sum, boys and girls in both programmes improved their mean achievement scores by a statistically significant amount, and neither programme showed differential effectiveness by gender.

Table 7.6: Mean achievement scores (and standard deviations) by gender, programme and time, and change in score

	JUMP (N=271)		IMPACT (N=238)	
	Girls	Boys	Girls	Boys
Sept. 2013	100.8 (13.7)	102.7 (13.6)	100.9 (15.9)	103.2 (15.0)
May 2014	108.2 (14.8)	109.3 (14.1)	106.5 (18.3)	108.2 (17.3)
Change	+7.4 (9.1)	+6.6 (8.6)	+5.6 (9.5)	+5.0 (10.1)

Results of initially low- and high-achieving pupils

This section analyses the results of two subsets of pupils:

- Low-achieving pupils, operationally defined as those whose DPMT score in September 2013 was more than one standard deviation (15 scale score points) *below* the national mean.
- High-achieving pupils, operationally defined as those whose DPMT score in September 2013 was more than one standard deviation *above* the national mean.

In the initial administration of the DPMT, 53 pupils from the core group achieved scores more than one standard deviation lower than the national mean, while 107 pupils achieved scores more than one standard deviation above the mean (Table 7.7). The 22 low-achieving JUMP pupils had a mean score of 80, and the 31 IMPACT pupils had a significantly lower mean score of 77 (*t*(49)=2.64, *p*=.011). The 56 high-achieving JUMP pupils obtained a mean of 122, very similar to that of 123 obtained by the 51 high-achieving IMPACT pupils. Thus, both groups were closely matched on numbers and mean scores of high-achieving pupils, but IMPACT had slightly more low achievers, with a slightly lower mean score (80 versus 77).

Looking at how the same pupils performed in May 2014, low-achieving pupils in both programmes improved by about five score points (one third of a standard deviation). These increases were significant in both programmes ([JUMP: *t*(21)=-2.81, *p*=.011], [IMPACT: *t*(30)=-3.87, *p*=.001]). The difference across programmes between scores of low-achieving pupils was no longer statistically significant at the end of the year (*t*(51)=1.20, *p*=.236).

Over the course of the evaluation, the scores of high-achievers also improved (significantly), by two points in JUMP and three points in IMPACT ([JUMP: *t*(55)=-2.50, *p*=.015], [IMPACT: *t*(50)=-2.56, *p*=.013]). However, the results of high-achieving JUMP pupils did not differ significantly from those of high-achieving IMPACT pupils in September (*t*(105)=1.10, *p*=.274) or May (*t*(86)=1.38, *p*=.173).

The number of cases is very small when achievement and programme are further split by gender (ranging from only 10 girls in the JUMP low-achieving group to 35 boys in the JUMP high-achieving group). Given the small numbers involved, tests of statistical significance were not deemed appropriate. However, the mean scores can still be described, once interpreted with caution (Table 7.7). Examining low-achieving pupils by gender, girls in both JUMP and IMPACT achieved lower mean scores than boys on both occasions. The JUMP gender gap remained almost identical (2.9 points at the start of the year and 2.8 points at the end), while the IMPACT gender gap increased slightly (from 2.3 points to 3.9 points). Examining high-achieving pupils by gender, girls in both programmes began in September with slightly lower mean scores than boys in September, but ended with slightly higher scores than boys in May. Among high achievers, the scores of IMPACT girls increased the most (4.1 points).

Table 7.7: Mean achievement scores (and standard deviations) for pupils whose baseline scores were more than one standard deviation below/above the mean, overall and by gender, programme and test time

		JUMP			IMPACT		
		Sept ’13	May ’14	Diff	Sept ’13	May ’14	Diff
1 SD below	Total (N=53)	80.0 (3.1)	85.5 (9.8)	+5.5 (9.3)	76.8 (5.5)	82.1 (10.3)	+5.3 (7.7)
	Girls (N=25)	78.4 (3.7)	84.0 (9.4)	+5.6 (8.4)	75.6 (5.8)	80.1 (9.3)	+4.5 (6.1)
	Boys (N=28)	81.2 (1.7)	86.7 (10.3)	+5.5 (10.3)	77.9 (5.2)	84.0 (11.2)	+6.1 (9.0)
1 SD above	Total (N=107)	122.1 (6.1)	124.2 (6.5)	+2.1 (6.4)	123.4 (6.1)	126.4 (9.6)	+3.0 (8.5)
	Girls (N=43)	121.9 (5.0)	124.4 (7.3)	+2.6 (7.1)	122.9 (5.6)	127.0 (9.3)	+4.1 (8.4)
	Boys (N=64)	122.2 (6.7)	124.1 (6.0)	+1.9 (6.0)	123.7 (6.6)	125.9 (10.0)	+2.2 (8.6)

Observation type and pupil achievement

As described in Chapter 4, mathematics lessons were observed “live” in 11 classes (on two occasions) and recorded in 15 classes (also on two occasions). Many teachers indicated that they were particularly nervous about being recorded while teaching, while some also requested copies of the recorded lessons. On the basis that either the act of being recorded or the effects of watching a recorded lesson might affect teaching behaviour (e.g., improved teaching due to reflecting on observed teaching behaviour, extra effort due to being recorded, poorer quality teaching due to the stress of being recorded), pupil results were analysed by observation type.

JUMP pupils in recorded classes had a higher mean increase in scores than JUMP pupils in non-recorded classes (eight score points as opposed to six score points). For IMPACT pupils, the mean baseline score was higher among pupils in non-recorded classes, but there was almost no difference in mean increases between those who were recorded and those who were not (Table 7.8). Significant increases in mean scores were evident in both recorded and non-recorded groups in JUMP ([recorded: *t*(121)=-11.07, *p*<.001], [non-recorded: *t*(148)=-7.75, *p*<.001]) and IMPACT ([recorded: *t*(134)=-5.60, *p*<.001], [non-recorded: *t*(102)=-6.38, *p*<.001]). However, the difference in the size of the increase *between* recorded and non-recorded groups was negligible for IMPACT, and not significant for JUMP on either occasion ([September: *t*(269)=-.37, *p*=.713], [May: *t*(269)=1.21, *p*=.227]). In other words, observation type did not appear to have a significant effect on pupils’ achievement results.

Table 7.8: Mean achievement scores and standard deviations, by observation type, programme, and test time

	Recorded observation		Non-recorded observation	
	JUMP (N=122)	IMPACT (N=135)	JUMP (N=149)	IMPACT (N=103)
Sept 2013	101.6 (14.6)	100.0 (15.5)	102.2 (12.9)	105.0 (14.8)
May 2014	110.0 (14.5)	105.0 (18.2)	107.9 (14.3)	110.7 (16.8)
Change	+8.4	+5.0	+5.7	+5.7

Teacher variables and pupil achievement

In this section, pupils' achievement gains are analysed vis-à-vis four teacher characteristics:

- Mathematical knowledge for teaching (MKT).
- Mathematical quality of instruction (MQI).
- CPD uptake.
- Programme adherence.

Mathematical knowledge for teaching

The shortened Mathematical Knowledge for Teaching Questionnaire (MKTQ-S) was completed by 31 teachers in September 2013, and by 26 teachers in May 2014. Data from the MKTQ-S were used for two purposes, bulleted below and described in the next sections.

- to assess whether teachers' mathematical knowledge for teaching changed during the year, using the results of a core group of teachers who sat the test on both occasions.
- to assess whether pupils' achievement was related to their teacher score on the MKTQ-S, using baseline teacher results where available.

Overall scores on the MKTQ-S

The core group that completed the MKTQ-S on both occasions comprised 23 teachers: 14 from JUMP (representing 11 classes) and nine from IMPACT (representing nine classes). The small numbers reflect a slight drop in response rate at the end of the evaluation, and the difficulty in "matching" data for teachers who were job-sharing, on leave, or working as learning support teachers in a number of classes. Scores are reported as the number of questions answered correctly, relative to the Irish norm group of teachers (Delaney, 2012).

In September 2013, the MKTQ-S scores of core group teachers in JUMP and IMPACT differed by less than one question answered correctly (Table 7.9). Both figures are slightly higher than the mean score achieved on MKTQ-S questions by a previous sample of 500 Irish teachers, who were administered the whole MKTQ (S. Delaney, personal communication, August 2013). This may reflect the self-selection of teachers interested in mathematics, as well the shorter length of the MKTQ-S.[4]

In May 2014, teachers in both programmes answered an average of roughly four and a half more questions correctly than did the Irish norm group. This is equivalent to a 10% advantage on the MKTQ-S for teachers in the evaluation, relative to the norm group. However, there was considerable individual variation in both groups. The number of questions answered correctly by two teachers (one in each programme) decreased by five, while questions answered correctly increased by nine for two JUMP teachers and by 17 for one IMPACT teacher. However, in most cases, the difference in start- and end-of-year scores ranged between a decrease of one and an increase of four correct answers.

Table 7.9: Difference in the number of MKTQ-S questions answered correctly by teachers in the core group (by programme and time) versus Delaney's Irish norm group

JUMP (N=14)		IMPACT (N=9)	
Time 1	Time 2	Time 1	Time 2
+1.4	+4.4	+2.1	+4.3

[4] Although the MKTQ-S is a halved form of the MKTQ, it usually took at least 40 minutes to complete. As such, it is possible that fatigue may depress scores for the full version.

Teachers' MKTQ-S and pupil achievement

Teacher MKTQ-S scores were applied to classes, using teachers' results from September where possible, to provide a baseline measure of their mathematical knowledge for teaching. For three teachers for whom September data were not available, May results were used. The mean score for JUMP classes was two correct answers fewer than for IMPACT classes.

When mean class MKTQ-S scores were correlated with gains in pupil achievement (Pearson, two-tailed), a small but statistically significant *negative* correlation was found for JUMP (*r*=-.17, N=271, *p*=.005) (Table 7.10). That is, JUMP classes with lower mean MKTQ-S scores tended to have slightly *higher* gains in pupil achievement over the course of the evaluation. However, this correlation represents a very small percentage of variance (2.9%). There was not a significant correlation for IMPACT. Correlations between mean class MKTQ-S scores and end-of-year pupil achievement scores were not significant for either programme.

Table 7.10: Correlation between baseline class MKTQ-S score and: (a) pupil difference score; (b) pupil achievement score in May 2014

	JUMP (N=271)	IMPACT (N=238)
r for pupil difference score and class MKTQ-S score	- 0.17**	0.03
r for pupil May 2014 DPMT score and class MKTQ-S score	0.04	0.10

** Significant at 0.01 level.

Mathematical quality of instruction (MQI)

Recorded observations were conducted at two points during the year for 15 classes, and were rated on the MQI by two observers on each occasion (see Chapter 4). As some teachers indicated that they were nervous and/or taught an atypical lesson for the first observations, ratings from the second observations were considered more reflective of teachers' general practice. Therefore, drawing on the two observers' overall MQI ratings from the second observation *only*, the mean rating was 1.9 for JUMP and 2.1 for IMPACT. As the MQI is a three-point scale (where 1 equals low quality and 3 equals high quality), both means might be considered to reflect medium quality levels.

Class MQI ratings were correlated (Pearson, two-tailed) with gains in pupil achievement and with pupils' May scores (Table 7.11). Neither correlation was significant for JUMP pupils, but both were significant for IMPACT pupils ([MQI and gains in pupil achievement: *r*=.25, N=135, *p*=.004, corresponding to 6% of variance], [MQI and May pupil achievement: *r*=.40, N=135, *p*<.001, corresponding to 16% of variance]). Thus, gains and overall achievement scores were higher for pupils taught by IMPACT teachers who were rated as displaying higher quality of instruction.

Table 7.11: Correlation between class MQI rating from second observations, and: (a) gains in pupil achievement; (b) pupil achievement score in May 2014

	JUMP (N=122)	IMPACT (N=135)
r for pupil difference score and MQI rating	-0.04	0.25**
r for pupil May 2014 DPMT score and MQI rating	0.12	0.40**

**Significant at 0.01 level.

CPD uptake

Linking pupil achievement and teacher uptake of CPD is somewhat problematic. There are qualitative differences between attending the initial training day, viewing it online at a later date, attending some or all of a webinar, or doing all of the preceding. Thus, CPD attendance cannot simply be summed and correlated with achievement, meaning teacher uptake of CPD falls into multiple (not directly comparable) categories. Given a small number of teachers, already divided by programme, most categories of CPD fall below the point where analyses of statistical significance are appropriate. Therefore, this section only examines teacher attendance at the initial training day. Attendance is defined as physical attendance, and does not include the small number who later watched recordings of part or all of the initial sessions.

Teachers from nine JUMP and nine IMPACT classes attended the initial training day.[5] JUMP pupils whose teachers attended initial training showed an increase of 7.7 points in mean scores from September 2013 to May 2014, while JUMP pupils whose teachers had not attended showed an increase of 5.0 points. The difference in gains between these groups was statistically significant (*t*(269)=2.34, *p*=.020). Amongst IMPACT pupils, the average increase in scores was 4.9 points for pupils whose teachers attended, and 5.7 points for those whose teachers did not. This difference was not statistically significant (*t*(236)=-0.62, *p*=.539).

Programme adherence

As with MQI ratings, programme adherence ratings from the second observations were thought to be more reliable than those from the first observations. No significant correlations were found between programme adherence and gains in pupil achievement (Table 7.12). Programme adherence was significantly positively correlated with IMPACT pupils' achievement scores in May (*r*=.17, N=225, *p*=.009).

Table 7.12: Correlation between class adherence rating from second observations, and: (a) pupil difference score; (b) pupil achievement score in May 2014

	JUMP (N=271)	IMPACT (N=225)
r for pupil difference score and adherence	-0.09	0.04
r for pupil May 2014 DPMT score and adherence	0.05	0.17**

**Significant at 0.01 level.

The sample of teachers is too small to warrant reporting the correlation values for adherence and other teacher variables, e.g., MKT baseline score, MKT difference score, and MQI rating (second observation). However, no strong associations were apparent between programme adherence and any of these variables.

Pupil attitudes and pupil achievement

This section examines the relationship between pupils' attitudes to mathematics (as reported in Chapter 5) and their mathematical achievement. Findings are grouped as relating to pupils':

- *general attitudes* to mathematics.
- *confidence* in their own mathematical competence.
- reports of their *teachers' behaviour* in mathematics class.
- *habits* when learning and practising mathematics.

[5] In a tenth IMPACT class, the teacher who attended initial training subsequently went on leave and was replaced by a new class teacher, who had not attended initial training.

General attitudes to mathematics

At the start and end of the year, pupils were asked to agree or disagree (a little or a lot) with the statements that they *liked mathematics* and that they *wished they didn't study mathematics.*

In September, pupils' liking of mathematics was significantly positively correlated with achievement in JUMP (r_s=.17, N=255, *p*=.007), but not IMPACT (r_s=.12, N=221, *p*=.080). Similarly, there was a significant negative correlation between pupil achievement and wishing they didn't study mathematics for JUMP (r_s=-.21, N=266, *p*<.001), but not IMPACT (r_s=-.02, N=227, *p*=.714). However, in May, this pattern was reversed. For IMPACT, achievement was significantly positively correlated with liking mathematics (r_s=.16, N=231, *p*=.013), but this was not the case for JUMP (r_s=.010, N=268, *p*=.095). Achievement in both groups was significantly negatively correlated with not wanting to study maths ([JUMP: r_s=-.15, N=270, *p*=.02], [IMPACT: r_s=.23, N=237, *p*<.001]) (Table 7.13). Thus, during the evaluation, the association between positive attitudes to mathematics and achieving high scores decreased in JUMP, but increased in IMPACT.

Table 7.13: Correlation between pupils' mean achievement scores and two variables indicating general attitude to mathematics, by programme and test time

	JUMP		IMPACT	
	Sep. '13	May '14	Sep. '13	May '14
r_s for achievement score and *I like maths*	0.17**	0.01	0.12	0.16
r_s for achievement score and *I wish I didn't study maths*	-0.21**	-0.15*	-0.02	-0.23**

*Significant at 0.05 level **Significant at 0.01 level

The decreased association between attitude and achievement in JUMP is explained in part by the fact that, in May, JUMP pupils who "disagreed a little" that they liked maths had generally high achievement scores (mean scaled score 112, range 94-133, N=29), as did those who "agreed a little" that they didn't want to study maths (mean score 110, range 84-136, N=75) (Table 7.14). These groups of pupils also had high mean increases in achievement, averaging to nine score points in both cases. Therefore, some JUMP pupils whose achievement scores *and* gains were high at the end of the evaluation had more negative attitudes to mathematics than might have been expected.

Table 7.14: Mean maths achievement scores for pupils who agreed to various extents with two statements concerning attitude to mathematics, by programme and test time

		JUMP		IMPACT	
		Sep. '13	May '14	Sep. '13	May '14
"I like maths"	Agree a lot	104.6	109.9	104.8	109.7
	Agree a little	102.8	108.9	103.7	109.6
	Disagree a little	97.4	111.7	101.6	105.5
	Disagree a lot	99.4	103.8	99.4	100.8
"I wish I didn't study maths"	Agree a lot	98.3	102.9	102.8	99.8
	Agree a little	100.4	110.1	101.3	105.3
	Disagree a little	103.5	109.8	104.2	110.6
	Disagree a lot	105.5	110.4	102.8	110.7

Confidence in mathematical ability

At the start and end of the year, pupils were asked to agree or disagree (a lot or a little) with statements relating to mathematical confidence and anxiety: including that that they were good at maths, that they worried about not being able to answer questions in maths class, and that they thought everyone could be good at maths.

For pupils in both programmes, on both occasions, positive correlations between belief in being good at maths and achievement score were significant, although slightly larger in September ([JUMP: r_s=.29, N=263, p<.001], [IMPACT: r_s=.33, N=231, p<.001]) than in May ([JUMP: r_s=.22, N=271, p<.001] [IMPACT: r_s=.27, N=236, p<.001]) (Table 7.15).

There was also a significant negative correlation between achievement scores and the extent to which pupils agreed with the statement *I worry I won't be able to answer questions in maths class*, in both programmes and on both occasions. The size of the correlation decreased slightly for JUMP, but increased for IMPACT, from September ([JUMP: r_s=-.31, N=264, p<.001], [IMPACT: r_s=-.17, N=229, p=.01]) to May ([JUMP: r_s=-.21, N=266, p<.001], [IMPACT: r_s=-.35, N=237, p<.001]). In other words, for IMPACT, the association between low achievement and pupil anxiety levels about being asked questions increased over the course of the year.

Table 7.15: Correlation between pupils' mean achievement scores and two variables indicating confidence in their own mathematical competence, by programme and test time

	JUMP		IMPACT	
	Sep. '13	May '14	Sep. '13	May '14
r_S for DPMT score and pupils agreeing *I am good at maths*	0.29**	0.22**	.33**	0.27**
r_S for DPMT score and *I worry I won't be able to answer questions in maths class*	-0.31**	-0.21**	-0.17*	-0.35**

*Significant at 0.05 level **Significant at 0.01 level

The belief that everyone can be good at maths was not significantly correlated with achievement for pupils in either programme, at either test time. However, among JUMP pupils, there was a significant positive correlation at the start of the year between how much they believed *they* were good at mathematics and how much they believed *everyone* could be good at maths (r_s=.23, N=251, p<.001). At the end of the year, the correlation was not significant (r_s=.10, N=267, p=.107) (Table 7.16). For IMPACT pupils, there was no significant correlation, on either occasion, between belief in their own mathematical ability and belief in the potential of everyone to be good at mathematics.

Table 7.16: Correlation between pupils' belief that *everyone could be good at maths* and (a) achievement scores; (b) belief that *they are good at maths*, by programme and test time

	JUMP		IMPACT	
	Sep. '13	May '14	Sep. '13	May '14
r_S for achievement score and pupils agreeing that *everyone can be good at maths*	-0.04	0.02	-0.04	-0.08
r_S for pupils believing *I am good at maths* and that *everyone can be good at maths*	0.23**	0.08	0.10	0.04

**Significant at 0.01 level

Teachers' behaviour in mathematics class

Pupils were asked to agree or disagree (a lot or a little) with statements about what their teacher did in their mathematics lessons. Those that were related to achievement are shown in Table 7.17. In September, pupils' agreement that their teacher *always asked if they understood* was positively correlated with achievement for both programmes ([JUMP: r_s=.13, N=266, p=.043], [IMPACT: r_s=.18, N=227, p=.005]). However, this correlation was not significant for either programme in May.

JUMP pupils' agreement that their teacher *got them to practice lots of examples* was *negatively* correlated with achievement at both test times, with correlation size increasing on the second occasion ([September: r_s=-.14, N=266, p=.020], [May: r_s=-.24, N=267, p<.001]). The correlation was not significant for IMPACT pupils at either test time.

At the outset, pupils' beliefs that their teacher *gave them fun things to do* and *let them play games* were not significantly correlated with achievement for either programme. However, at the end of the year, agreement that the teacher gave them fun things to do was *negatively* correlated with achievement in both groups ([JUMP: r_s=-.17, N=270, p=.006], [IMPACT: r_s=-.13, N=236, p=.049]). For JUMP pupils, agreement that the teacher let them play games was also negatively correlated with achievement (r_s=-.20, N=271, p=.001), though this correlation was not significant in IMPACT.

At the end of the year, there was a small negative correlation between JUMP pupils' achievement and their agreement that their teacher *often praised them* (r_s=-0.14, N=266, p=.025). This correlation was not significant for IMPACT.

Table 7.17: Correlation between pupils' achievement scores and their agreement with various statements about their teachers' behaviour in mathematics class

	JUMP		IMPACT	
	Sep. '13	May '14	Sep. '13	May '14
"My teacher always asks do we understand stuff."	0.13*	-0.06	0.18**	0.10
"My teacher often praises me."	0.04	-0.14*	0.07	-0.07
"My teacher gets me to practice lots of examples."	-0.14*	-0.24**	0.07	-0.02
"My teacher gives us fun things to do."	0.50	-0.17**	0.13	-0.13*
"My teacher lets us play games."	0.01	-0.20**	0.07	0.07

*Significant at 0.05 level **Significant at 0.01 level

Habits when learning mathematics

Pupils were asked to indicate how frequently they used various learning strategies for mathematics (*every class*, *most classes*, *some classes*, or *hardly ever*). Table 7.18 shows those significantly correlated with achievement.

In September, frequency of repeating examples was not significantly correlated with achievement for either group, but in May there was a significant *negative* correlation for both JUMP (r_s=-.20, N=265, p=.001) and IMPACT (r_s=-0.18, N=234, p=.006). That is, at the end of the year, pupils who reported frequently repeating examples were likely to have lower achievement scores than those who did not.

At the start of the year, there was a significant negative correlation between achievement and how often pupils worked with their classmates to solve a problem for JUMP (r_s=-0.18, N=258, p=.005), but not IMPACT. At the end of the year, however, a negative correlation was significant for IMPACT (r_s=-.16, N=229, p=.019), but not JUMP. For the item asking how often pupils worked out a sum in their heads, a significant negative correlation with achievement was observed for IMPACT pupils in May (r_s=-.15, N=233, *p*=.020).

Table 7.18: Correlation between pupils' achievement scores and the frequency with which pupils reported engaging in various learning strategies for mathematics

	JUMP		IMPACT	
	Time 1	Time 2	Time 1	Time 2
"When we do new things, I learn as much as I can by heart."	0.16**	0.07	0.04	-0.12
"I go through examples again and again to help me remember them."	-0.06	-0.20**	0.01	-0.18**
"I work with my classmates to solve a problem."	-0.18**	-0.11	-0.07	-0.16*
"I work out a sum in my head."	0.04	-0.05	0.03	-0.15*

*Significant at 0.05 level **Significant at 0.01 level

Summary

Pupils' achievement improved significantly in *both* programmes over the course of the evaluation. On average, JUMP pupils improved slightly more than IMPACT pupils (seven scale score points as opposed to five), but this difference was not statistically significant.

When results were broken down by strand and process, there were few programme-based differences. However, in May, JUMP pupils outperformed IMPACT pupils on the Data strand and on the process of Integrating and Connecting. These differences are largely attributable to pupils' performance on a few linked items relating to pictogram interpretation.

Gender differences were small, with boys slightly, but not significantly, outperforming girls across the whole sample and within the two programmes on both occasions. Achievement gains were significant for both girls and boys in the two groups.

In both programmes, pupils who achieved low scores in the September tests gained about five points by the May tests, on average. Pupils who achieved high scores at the start of the evaluation gained about two points in JUMP and three in IMPACT.

Teachers' scores on the MKTQ-S improved slightly in both JUMP and IMPACT during the evaluation. However, teachers' (initial) Mathematical Knowledge for Teaching scores were not related to pupils' end-of-year achievement scores. JUMP teachers' MKT scores were negatively correlated with their pupils' *gains* in achievement, indicating that teachers with lower MKT scores saw bigger improvements in their pupils' performance over the year than did teachers with higher MKT scores.

Mathematical Quality of Instruction (as rated by subject matter experts) was unrelated to pupil achievement in JUMP. In IMPACT, however, there was a significant positive correlation between MQI and both pupils' end-of-year scores and gains in achievement. JUMP pupils whose teachers had attended the initial training day showed slightly, but significantly higher gains those whose teachers did not attend. No attendance effect was apparent for IMPACT. In contrast, teachers' adherence to programme ethos was not linked to achievement in JUMP, but was positively correlated with pupils' end-of-year scores in IMPACT.

At the start of the year, pupil achievement was significantly correlated with having a positive general attitude towards mathematics in JUMP, but not IMPACT. However, at the end of the year, this pattern was reversed, partly because high-achieving (and high-gaining) JUMP pupils now expressed slight dislike for mathematics. High confidence and low anxiety were correlated with achievement for pupils in both programmes, both at the start and end of the evaluation, although this pattern was stronger in IMPACT than JUMP. In particular, the relationship between low achievement and anxiety about being asked questions in class was more pronounced in IMPACT pupils at the end of the year.

Few teacher behaviours (as reported by pupils) were positively correlated with achievement. In May, achievement in both programmes was *negatively* correlated with teachers frequently "giving pupils fun things to do". In JUMP, it was also negatively correlated with teachers letting pupils play games, getting them to practice a lot of examples, and praising them.

In May, for both programmes, achievement was negatively correlated with pupils reporting frequent use of repeated examples in their own mathematics practice. In IMPACT, it was also negatively correlated with the frequency of pupils working with classmates, and working out sums in their heads.

Overall, the achievement results indicate significant improvements in both programmes, but do little to distinguish the programmes' effects *from one another*.

Note on significance levels

When more than two mean scores are simultaneously compared, there is an increased probability of what is called a "Type 1 error", or a false positive. That is, the more comparisons are made, the greater the likelihood that the groups will differ on at least one comparison.

The Dunn-Bonferroni procedure (see for instance Kirk, 1968) can be used to control for this possibility, by calculating an *adjusted* significance level. The significance level is divided by the number of related t-tests carried out, with the resulting figure divided by two for a two-tailed test. For instance, if the desired significance level for a single t-test is .05 (that is, a one in twenty chance of a Type 1 error), and eight related t-tests are carried out, the adjusted significance level is .003. This is obtained as follows: [$\frac{0.5}{8}$ = 0.006], then [$\frac{0.006}{2}$ = .003]. In this instance, each of the eight t-tests would only be statistically significant if the *p* value was less than .003, not the original .05.

The data reported throughout this chapter draw on uncorrected significance levels, in part because corrections for multiple comparisons, coupled with a relatively small sample, can lead to extremely conservative interpretation of significance. When the Dunn-Bonferroni adjustment is applied, four results move from significant to non-significant:

1. JUMP and IMPACT pupils' overall September scores are not significantly different from the population mean (adjusted significance level=.003).
2. The difference between low-achieving pupils in JUMP and IMPACT at the start of the year is not significant (adjusted significance level=.006).
3. The gains made by low-achieving JUMP pupils during the year are not significant (adjusted significance level=.006).
   However, the difference in *p* values for JUMP and IMPACT in this regard is due partly to the smaller number of JUMP than IMPACT pupils in the low-achieving category.
4. The gains made by high-achieving pupils in both programmes are not statistically significant (adjusted significance level=.006).