Chapter 4
Classroom observations

As outlined in Chapter 2, classroom observations were conducted for two main reasons. First, they provided an indication of the extent to which programme principles were reflected in teachers' classroom practices. Second, they allowed analyses of differences, if any, between the mathematical quality of instruction experienced by pupils in each programme.

To measure *adherence* to programme principles, a tailored observation schedule was developed (incorporating specific adherence indicators and more global ratings of adherence). To measure the *quality* of mathematical instruction, recorded lessons were rated on a modified version of the Mathematical Quality of Instruction (MQI) instrument (Learning Mathematics for Teaching Project, 2011) and described using qualitative lesson reports.

This chapter is divided into five sections, the first of which explains the types of observation data collected. The second section describes the curriculum areas covered in the lessons observed. The third section describes the extent of adherence to programme methods and principles, and the fourth analyses the quality of mathematical instruction in the lessons observed. The fifth offers some general comments on the observed lessons, based on the overall impressions of observers.

Types of observation data

The content of this chapter is based on 52 mathematics lessons that were observed as part of the evaluation. Twenty-six classes were observed between December 2013 and January 2014, and again in May 2014. On both occasions, trained observers observed lessons "live" in 11 classes, while lessons were recorded for later analysis in 15 classes. In the first set of observations, average lesson duration for both programmes was 44 minutes and ranged between 30 and 60 minutes. In the second set, average duration was again 44 minutes for JUMP lessons (range 23 to 60 minutes), while IMPACT lessons averaged 46 minutes (range 24 to 73 minutes).

Three main types of observation data were collected: those from a tailored observation schedule, ratings from an abridged MQI, and a summary lesson report.

Tailored observation schedule

All 52 lessons (i.e., both live and recorded) were observed using a tailored observation schedule. The activities targeted in the schedule were those expected to produce maximum difference between the two programmes, were teachers to adhere closely to programme principles. Thus, high levels of adherence to JUMP principles and methods were expected to manifest in lessons emphasising high levels of teacher-led instruction, low incidence of group work, frequent use of workbooks, bonus questions, memorisation and repeated practice of procedures to acquire mastery of mathematical concepts and skills. In contrast, close adherence to IMPACT would be likely to lead to lessons that emphasised pupil-led discussion, group work, and frequent use of

concrete and pictorial materials before the introduction of the abstract. IMPACT teachers were expected to encourage collaborative problem-solving and guided discovery.[1]

The observation schedule comprised a mixture of rating types. Some variables would be expected to be observed very regularly in lessons, meaning that frequency was relevant. Examples include teacher-led instruction, or pupils engaged *on task*. For such variables, the observers indicated what percentage of each of a series of five-minute lesson segments was spent on each activity. The percentages of time within each segment were summed to produce an overall estimate of the percentage of observed class time for which each activity was observed.

The main activities rated in this way were:

- teacher providing instruction or information to pupils.
- pupils answering teacher questions.
- pupils asking questions or discussing maths (with others, or with teacher).
- pupils working as individuals (with or without help from teacher).
- pairs/small groups working on maths.
- time spent *on task*.

The percentages of time spent on the above activities do not sum to 100% as many overlap. For example, a teacher may be providing instruction while pupils are focused on task.

Another type of rating required the observers to draw on their professional experience as teachers to provide a more *global quality* rating. Examples of this type of rating include the extent to which differentiated teaching practices were evident during the lesson, and to which pupils seemed engaged. Finally, some ratings related to a global estimate of the amount of time spent on an activity. For example, observers provided an overall estimate of the percentage of class time pupils spent working from their workbooks/textbooks.

MQI ratings

All recorded lessons were observed by two subject matter experts (SMEs) and rated on a modified version of the MQI. In total, 60 sets of ratings were generated (30 lessons, each rated by two people). Lessons were rated as low, medium, or high on four dimensions:

- richness of the mathematics.
- working with pupils and mathematics.
- errors and imprecision.
- pupil participation in meaning-making and reasoning.

The dimension ratings were themselves based on a number of more specific indicators. For example, five items fed into the overall richness of the mathematics rating, while errors and imprecision and pupil participation were each based on three specific indicators, and working with pupils was based on two specific indicators.

[1] Of course, these expectations are based on the *relative*, not absolute, principles of JUMP and IMPACT. As outlined in Chapter 1, JUMP also promotes some guided discovery, although its definition of the term is different to that used generally. Equally, IMPACT promotes some scaffolding of discovery activities.

In most cases, a rating of high was a positive rating (e.g., teachers might reasonably be pleased if the *richness of the mathematics* in their lesson was rated as high). However, for errors and imprecision, a low rating was indicative of good quality teaching, as it meant that the teacher displayed few mathematical errors or instances of imprecise language. In addition to the four main dimensions, each SME rated lessons on two global dimensions (Mathematical Quality of Instruction [MQI] and Mathematical Knowledge for Teaching [MKT]).

SMEs reviewed and rated each lesson separately. Once ratings were completed, the two sets of ratings were compared. Given the quite limited nature of a three-point scale, it was unsurprising that SMEs varied on some ratings. Nonetheless, most ratings were identical or within a point of each other (e.g., one SME assigned a medium rating while another assigned a high). For the few occasions where there was a marked difference (low versus high), the SMEs and project researchers jointly reviewed the ratings and the lesson, coming to an agreed rating.

Lesson Report

SMEs watched each recorded lesson on a number of occasions. On the first occasion, they took notes and wrote a short lesson report. These reports described the lesson's structure, main activities, the time spent on each activity, and the details of the types of materials used. As it was expected that (if teachers adhered to programme principles) JUMP teachers would make less use of manipulatives and real life materials in a discovery context than would the IMPACT teachers, the lesson report notes these elements. As well as describing the lesson, the reports were drawn on by the SMEs when later rating specific aspects of the lesson.

Curriculum areas covered

The strands and strand units covered in each observed lesson (live or recorded) were noted.

Strand

In the first set of observations, Number was the strand most commonly covered, featuring in seven JUMP and 11 IMPACT lessons (Table 4.1). The next most common strand was Measures (four JUMP lessons and one IMPACT lesson), followed by Algebra, and Shape and Space. No lessons in either group dealt with Data. In contrast, by the second set of observations, almost all IMPACT lessons (10 of the 13) covered a Shape and Space topic, as did four JUMP lessons. However, other JUMP lessons covered topics in the Number, Measures and Data strands.

The marked popularity of Number, then Shape and Space topics among IMPACT teachers was probably related to the content of IMPACT manuals. The first IMPACT manual received at the start of the academic year dealt only with two Number strand units. The second manual, received closer to Christmas, dealt only with Shape and Space. While IMPACT principles are intended to be generalisable across the whole PSMC, it is likely that teachers felt more comfortable being observed while teaching content covered by one of the manuals.

Table 4.1: Number of JUMP and IMPACT classes in which various PSMC strands were taught during observed lessons

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Number	7	3	11	1
Algebra	2	0	0	0
Shape and Space	0	4	1	10
Measures	4	3	1	2
Data	0	3	0	0

Strand unit

In the first observations, all Number strand units in JUMP classes related to either Place Value or Operations, the two Algebra lessons covered Number Patterns and Sequences, while the Measures strand units were Length (three lessons) and Time. In contrast, IMPACT mainly covered the Number strand units of Fractions (six lessons) and Decimals (four), with one lesson devoted to Operations. The only other strand units covered in IMPACT classes in the first set of observations were 3-D shapes (Shape and Space) and Length (Measures).

In the second set of observations, JUMP classes covered a wide variety of strand units. Three lessons covered Chance, two covered Fractions, with the following strand units covered in one of the observed lessons: Place Value, 2-D Shapes, 3-D Shapes, Symmetry, Capacity, Time, and Money. One lesson focused on coordinate points, which feature in JUMP but not the PSMC for Third class. IMPACT lessons were concentrated on the Shape and Space strand units of 2-D Shapes, 3-D Shapes, and Symmetry (10 lessons in total), with the three remaining lessons covering Fractions, Area, and Time.

Adherence to programme methods and principles

This section examines the extent to which lessons adhered to programme methods and principles, under seven broad headings:

- extent of teacher-led instruction (expected to be higher in JUMP classes).
- extent of pupil-led discussion (expected to be higher in IMPACT classes).
- pupil solo work (expected to be higher in JUMP classes) and group work (expected to be higher in IMPACT classes).
- types of materials used (workbooks expected to be more widely used in JUMP, and other materials expected to be more widely used in IMPACT).
- evidence of differentiated teaching practices (methods expected to differ by programme).
- learning styles (expected to differ by programme).
- use of assessment (expected to differ by programme).

The data presented are from the observation schedule, and relate to all 52 lessons observed.

Teacher-led instruction

Teacher-led instruction was operationally defined for observers as the teacher leading classroom instruction at whole-class level (including teacher-led question and answer sessions and work on the board, but excluding instruction of individual pupils or small groups while all pupils are carrying out solo or small-group work). For each of a series of five-minute time periods, observers rated the percentage of time spent in teacher-led instruction, with the percentages summed to gauge the percentage of total lesson time led by teachers.

For both sets of observations, a majority of class time in both programmes was taken up with teacher-led instruction (Table 4.2). However, the gap between programmes was smaller than might have been predicted during the first observations (averaging 67% of time in JUMP versus 58% in IMPACT lessons) and negligible during the second set of observations (57% in JUMP and 56% in IMPACT classes).

Question and answer sessions formed a subset of teacher-led instruction, and were defined as pupils answering questions *at whole-class level* (including listening to the teacher ask questions, waiting to answer questions [e.g., waving their hands] and actually answering questions). During the first observations, the two groups spent very similar percentages of class time on question and answer sessions (28% and 29% of JUMP and IMPACT lessons, respectively). By the second observations, the amount of time spent in teacher-led question and answer sessions had increased slightly for IMPACT lessons (to 36%), and increased noticeably for JUMP classes (to 41% of lesson time).

Table 4.2: Mean percentages of time spent on teacher-led instruction, questions and answers

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Teacher-led instruction	66.8	56.7	58.1	55.9
Teacher-led Q&A	27.9	41.1	29.4	36.5

Observers also gave a general estimate of the amount of time that pupils spent listening to the teacher talk to or question the class. The estimates broadly mirrored the summed percentages, showing that in both groups (but especially in JUMP), most pupils spent considerable time listening to the teacher (Table 4.3).

Table 4.3: Estimates of time spent by pupils listening to the teacher talk to or question the class

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Almost the entire lesson	1	2	0	0
Most of the time	7	5	6	5
About half the time	4	4	3	4
Some of the time	1	2	3	4
Hardly at all	0	0	1	0

Pupil-led discussion

Observers recorded the amount of time spent by pupils questioning or discussing mathematics, and listening to other pupils talk. They also rated the extent to which the classroom *climate* encouraged pupils to generate mathematical ideas and questions. As with teacher-led time, a series of five-minute ratings were summed to gauge the percentage of total lesson time spent in pupil-led discussion of mathematics. There were noticeable differences by programme, particularly during the first set of observations when only 7.0% of time in JUMP classes, but 31.4% of time in IMPACT classes, involved pupils questioning or discussing mathematics. By the second observations, this increased to 9.3% in JUMP, dropping to 25.1% in IMPACT classes.

Observers also gave a broad rating of the amount of time that pupils spent listening to other pupils talk. Although differences between programmes were less marked than for pupils discussing maths, the direction of difference was the same – i.e., pupils in JUMP classes were considered to have spent less time listening to other pupils talk than pupils in IMPACT classes (Table 4.4). For example, during the first set of observations, pupils in 10 JUMP and nine IMPACT classes spent either *some* or *hardly any* time listening to other pupils. By the second observations, this rose to 11 JUMP classes, while in IMPACT, it dropped to five classes.

Table 4.4: Estimates of time spent by pupils listening to other pupils talk

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Almost the entire lesson	0	0	0	0
Most of the time	1	1	1	3
About half the time	2	1	3	5
Some of the time	4	9	8	5
Hardly at all	6	2	1	0

During the first set of observations only five JUMP (nine IMPACT) lessons included at least five minutes of group or class discussion of a mathematical task or question, but this increased to eight JUMP lessons (and 10 in IMPACT) by the second observations. The slight increase in the amount of pupil discussion in JUMP classes is reflected in a slight change in how the observers rated classroom climate (Table 4.5). Initially, no JUMP class was rated as *definitely* encouraging of pupil ideas and questions, while four were rated as *probably* encouraging. By the second observations, three classes were *definitely* encouraging, and a further three were rated as *probably* encouraging. That aside, IMPACT classes were generally rated as more encouraging than JUMP classes of pupil ideas.

Table 4.5: Numbers of observed classes in which classroom climate was rated as encouraging pupil ideas and questions, to various degrees

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Definitely	0	3	3	4
Probably	4	3	2	4
Not sure	2	1	2	3
Not really	7	5	6	2
Not at all	0	1	0	0

Solo work and group work

Solo work was defined as pupils carrying out individual "seatwork" (e.g., working on workbooks or worksheets), including time spent by the teacher helping individual children. **Group work** was defined as time spent by pupils working in small groups or in pairs. There were large differences on these measures between the two programmes. On average, pupils in JUMP classes spent a relatively small amount of time in group work (14% and 11% for observations one and two, respectively), while IMPACT pupils spent close to half their lessons engaged in group work (42% and 43% for observations one and two, respectively) (Table 4.6). In a related vein, pupils in JUMP classes spent more time engaged in solo work – approximately 30% of time compared to 13% in IMPACT classes.

Table 4.6: Mean percentages of time spent on group and solo work

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Group work	13.9	11.0	41.5	43.2
Solo work	29.4	30.9	13.9	13.2

In addition to differing in the mean percentage of time spent in group work, JUMP classes were less likely to include *any* significant amount of group activity (Table 4.7). Across the 26 JUMP lessons observed, only nine included at least five minutes of group or pair work (compared to 21 IMPACT lessons). In contrast, 21 of the 26 JUMP lessons observed included at least five minutes of independent individual work, as did 18 IMPACT lessons.

Table 4.7: Number of lessons in which independent individual work or pair/group work was observed for at least five minutes

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Independent individual work	11	10	10	8
Independent pair or group work	4	5	11	10

Materials used

Observers were asked to indicate in which of a list of three types of activities (using textbooks/worksheets to answer questions, working with manipulatives, playing maths games) pupils were engaged for at least five minutes. On both occasions, the types of materials most widely used in JUMP were pupil questions in textbooks or workbooks, with less than half using manipulatives or maths games (Table 4.8). In contrast, manipulatives were used in most IMPACT lessons, with about half of classes also using maths games and questions in textbooks or workbooks.

Table 4.8: Number of classes that used particular materials/engaged in particular activities for at least five minutes

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Textbooks/worksheets	9	10	6	6
Manipulatives	4	5	11	12
Maths games	5	4	6	7

Drawing on broader information from the lesson reports, a variety of manipulatives were used by classes in both programmes (e.g., "real life" materials such as dice, counters, rulers/metre sticks, and chocolate boxes, and structured materials such as cubes, number cards, geoboards, and Dienes blocks).

Observers' estimates suggest that in addition to textbooks/worksheets being used in a larger number of JUMP classes, worksheets (or workbooks, in the JUMP programme) were used for longer periods of time. For example, during the first observations, textbooks/worksheets were used for 41-60% of lesson time in five JUMP but only two IMPACT classes (Table 4.9). During the second set of observations, eight JUMP classes used workbooks for at least 20% of class time, compared to only two IMPACT schools.[2]

[2] The amount of time in IMPACT schools may be slightly underestimated, as (unlike Table 4.8) it does not include time pupils spent copying questions from the textbook/the board and answering in their copybooks.

Table 4.9: Number of observed classes that spent within various percentage ranges of class time using worksheets/workbooks

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
61%+	0	0	0	0
41-60%	5	4	2	1
21-40%	1	4	1	1
1-20%	3	0	3	2
None	4	5	7	9

Differentiated teaching

As outlined in Chapter 1, bonus questions are proposed by the JUMP programme as a means of differentiating, since they should allow higher-achieving pupils to work independently while giving the teacher time to assist lower-achieving pupils by breaking down concepts and skills. The collaborative problem-solving methods advocated by IMPACT (including questioning and re-voicing techniques, and use of varied models) should also promote differentiated teaching, since they aim to allow pupils at various levels and with various learning styles to participate in developing solution methods. Therefore, observers provided an overall rating of the extent to which differentiated teaching practices were evident in the observed lesson, and, more specifically, if bonus questions and collaborative problem-solving were used in the lesson.

Clear use of differentiated teaching practices was observed in only a minority of lessons, most of which were IMPACT classes (Table 4.10). During the first set of observations, six JUMP lessons were rated as *not really* displaying differentiated practice, with an additional class where differentiated teaching was *not at all* present. By the second set of observations, five JUMP lessons were described as *not at all* showing differentiated practice, with a further four *not really* showing evidence of differentiation.

Table 4.10: Number of observed classes in which differentiated teaching was deemed present, to various degrees

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Definitely	0	1	2	5
Probably	3	2	2	0
Not sure	3	1	4	1
Not really	6	4	4	3
Not at all	1	5	1	4

Despite the relatively poor observer ratings for differentiation, bonus questions were used in nine JUMP classes during the first set of observations, falling to only five during the second set (Table 4.11). A teacher in one IMPACT class used a method similar to bonus questions during both observations. Thus, bonus questions were used less than might be anticipated in JUMP lessons, and, where used, were not always used in a manner that observers believed represented differentiated teaching practice.

Table 4.11: Number of observed classes in which bonus questions were given to pupils

	JUMP classes (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Bonus questions used	9	5	1	1

Collaborative problem-solving was not a feature of most JUMP classes during either set of observations, with observers rating it as *not really* or *not at all* manifested in a majority of classes (Table 4.12). However, observers also reported that it was *probably* or *definitely* a feature of only half of IMPACT classes.

Table 4.12: Number of observed classes in which collaborative problem-solving was deemed present, to various degrees

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Definitely	0	1	4	4
Probably	2	1	3	2
Not sure	1	2	0	1
Not really	5	4	4	0
Not at all	5	5	2	6

Learning styles

As noted in Chapter 1, a central tenet of the JUMP philosophy is that repetition and practice are key to learning. Therefore, observers were asked to report the extent to which these were present in the observed lessons. In the first set of observations, memorisation and repetition of procedures were either *definitely* or *probably* present in 11 JUMP classes, but also in four IMPACT classes (Table 4.13). However, during the second observations, JUMP classes were reasonably evenly split between those who showed some or no evidence of memorisation and repeat procedures.

Table 4.13: Number of observed classes in which memorisation and repeat procedures were deemed present, to various degrees

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
Definitely	7	5	4	2
Probably	4	2	0	2
Not sure	0	0	1	1
Not really	2	3	6	2
Not at all	0	3	1	6

Observers ranked the broad approaches of "simple direct instruction (explanation)", "instruction by a series of related questions" and "guided discovery (activities)", in order of their frequency of use in each lesson (Table 4.14). During both sets of observations, JUMP classes tended to use instruction by a series of questions or simple direct instruction, with guided discovery the least used strategy in almost all classes. In IMPACT classes, instruction by a series of questions was also the most common strategy, but in contrast to JUMP, guided discovery was the second most common strategy.

Table 4.14: Number of observed classes in which simple direct instruction, instruction by a series of questions, and guided discovery were rated as used with highest, medium, or lowest frequency

		JUMP (N=13)		IMPACT (N=13)	
		Time 1	Time 2	Time 1	Time 2
Simple direct instruction	Highest	6	6	4	1
	Medium	5	6	3	4
	Lowest	2	1	6	8
Series of questions	Highest	7	7	5	7
	Medium	6	6	6	5
	Lowest	0	0	2	1
Guided discovery	Highest	0	0	4	5
	Medium	2	1	4	4
	Lowest	11	12	5	4

Assessment

The observation schedule did not include assessment-related activities, due to the difficulty in pre-defining assessment practices in a manner that could be uniformly understood and rated. Instead, the SMEs' reports on the recorded classes provided data on how assessment was used in those 15 classes (30 lessons).

In both sets of observations, all recorded classes featured some form of continuous assessment. However, this was often limited to a teacher's monitoring of pupils' solo work. At other times, assessment was combined with oral review of material. In both these situations, it was usually the case that only *some* work of *some* pupils could be assessed. (That said, in two JUMP and two IMPACT classes within the recorded group, small class size or team teaching meant that *all* pupils could be informally assessed throughout the lessons).

In the first observations, more distinct forms of summative assessment were observed in three of the seven JUMP classes. In one class, pupils were given two worksheets as "quizzes". Each took approximately four minutes to complete, and all quizzes were collected by the teacher for correction. In the other two classes, approximately four minutes were spent checking answers at whole-class level, using the board. Two of the eight recorded IMPACT classes also devoted time to distinct forms of assessment. In one case, approximately 10 minutes were spent eliciting and discussing each pupil's solution to a problem (formative assessment). In the other case, two segments of approximately three minutes each were spent checking answers at whole-class level, using the board. Thus, programme-related patterns of assessment did not emerge strongly in the JUMP and IMPACT groups during the first observations.

In the second set of observations, there was even less evidence of distinct forms of assessment, although most JUMP and some IMPACT lessons included some oral review questions at whole-class level. One JUMP lesson also included a written review test, corrected verbally by the teacher at the start of the lesson, while another included time spent correcting workbook answers on the board. In general, however, assessment methods did not appear to differ notably by programme.

Global ratings of adherence to programme

For each lesson, observers rated on a scale of 1-10 how closely it adhered to the assigned programme. During the first set of observations, observers were asked to rate JUMP lessons for adherence to the selected lesson plan(s). For the second set of observations, observers provided two ratings of JUMP lessons – the degree of fidelity to the lesson plan and the degree of fidelity to JUMP principles. Ratings were split for the second observations because it was assumed that teachers might by that late stage in the year have become more comfortable with JUMP principles and proficient in JUMP methodologies. As such, they might be more likely to diverge from a set JUMP lesson plan, while still adhering to its general principles. Since the IMPACT programme did not contain lesson plans per se, global ratings for IMPACT were based on adherence to principles during both observations.

Most teachers in each group showed some levels of adherence to the relevant lesson plan or programme principles, but few showed very high levels. For JUMP, the average adherence rating to the lesson plan was 7.0 (out of 10) during the first set of observations, rising marginally to 7.2 for the second set (Table 4.15). With an average rating of 7.5, adherence to JUMP *principles* was slightly higher, although one lesson only received a rating of four out of 10. With IMPACT, average ratings for adherence to programme principles dropped between the first and second observations (from 7.2 to only 6.7). However, this can be attributed to one teacher, whose adherence ratings dropped from an initial rating of eight to two during the second observed lesson.

Table 4.15: Observers' overall ratings, on a scale of one (low) to 10 (high), of adherence to programmes (lesson plan and/or principles) in observed JUMP and IMPACT classes

	JUMP (N=13)			IMPACT (N=13)	
	Time 1	Time 2		Time 1	Time 2
	Lesson plan	Lesson plan	Principles	Principles	Principles
Mean rating	7.0	7.5	7.2	7.2	6.7
Range	5-9	5-9	4-9	5-9	2-9

Quality of instruction

To rate the *quality* of instruction in observed classes, the observation schedule (applied to all 52 observed lessons) included measures of pupil engagement and pupil understanding, while the subset of lessons which were recorded were also rated using a modified version of the MQI.

Pupil engagement

Three ratings of pupil engagement with the lesson were provided. First, observers indicated if they felt that *most or all* pupils were engaged with the lesson. Second, they estimated the proportion of pupils likely to have had a good understanding of the lesson by the end of class, and third, they rated the amount of time pupils spent "on task". Time on task is generally taken as meaning the amount of time in which pupils were actively engaged with their school work. In this instance it included pupils actively working on a task, paying attention to a teacher or a classmate, and the time taken to set up an activity *efficiently*.

Generally, pupil engagement with the lesson was high. Of the 26 JUMP lessons observed, pupils were rated as *definitely* or *probably* engaged with the lesson in all but one instance (Table 4.16). Engagement was also generally high in IMPACT classes, although in four lessons, observers were unsure of pupil engagement levels, and in one case pupils were rated as *not really* engaged.

Observers were slightly less positive when rating pupil understanding of lesson content. In the first observations, none of the 13 JUMP classes were rated as having all pupils possess a *good* understanding of the topic by the end of the lesson. However, in 10 classes, observers considered that over half the pupils had a good understanding. During the second observations, three observers felt that the entire class developed a good understanding of the lesson topic, but in four classes no more than about half of pupils were believed to have such an understanding. IMPACT classes showed a broadly similar pattern in that in most cases *over half* of pupils were believed to have a good understanding of the topic.

Table 4.16: Number of observed classes in which most or all pupils were considered engaged to various degrees, and in which various proportions of pupils had a good understanding of lesson content

		JUMP (N=13)		IMPACT (N=13)	
		Time 1	Time 2	Time 1	Time 2
Most/all pupils engaged with lesson	Definitely	4	5	5	4
	Probably	9	7	5	7
	Not sure	0	1	2	2
	Not really	0	0	1	0
	Not at all	0	0	0	0
Pupils with good understanding of content	All of them	0	3	3	1
	Over half	10	6	5	8
	About half	2	3	4	3
	Less than half	1	1	0	0
	None	0	0	0	0

Time on task was assessed in two ways, both of which suggested that for the vast majority of lesson time, pupils were generally on task. First, based on summed estimates from a series of five-minute lesson segments, JUMP pupils were on task 96.3% of the time during the first series of observations, and 94.3% of the time for the second set. For IMPACT, the equivalent data were 90.0% and 91.9%, respectively. Second, at the end of each lesson, observers provided a general estimate of how much time they felt was on task. Five options were presented, ranging from *less than 25%* to *over 90%* of time, but only the two highest were used by observers (Table 4.17). During both observations, a sizeable majority of JUMP classes were rated as spending over 90% of time on task. Ratings were also quite positive for IMPACT, although during the first set of observations six of the 13 lessons were rated as 76-90% on task.

Table 4.17: Number of classes in which percentage of class time spent on task was estimated to fall within the 91-100% or 76-90% ranges

	JUMP (N=13)		IMPACT (N=13)	
	Time 1	Time 2	Time 1	Time 2
91%+	9	10	7	9
76-90%	4	3	6	4

Mathematical Quality of Instruction (MQI)

The MQI (Learning Mathematics for Teaching Project, 2011) was used to rate all recorded observations on four broad dimensions (richness of mathematics, working with pupils and mathematics, teacher errors and imprecision, and pupil participation in meaning-making). Each of 15 class groups was recorded on two occasions, and rated by two SMEs, making a total of 60 sets of MQI ratings for the 30 individual lessons recorded.

The remaining tables in this chapter show SME ratings by programme and by observation. As will be seen, almost every rating improved between the first and second observations. This may be attributable to participation in the programmes improving quality of instruction, or to teachers generally reflecting more on their teaching practice as a result of participation in an evaluation. However, many teachers informally indicated that they were more comfortable in front of the camera on the second occasion, and less constrained by nervousness. It is therefore also possible that the data in subsequent tables may reflect this rather than any programme or evaluation effects.

Richness of the mathematics

Instruction was defined as featuring rich mathematics if it was focused on the meaning of mathematical facts and procedures, and/or deeply engaged with mathematical practices and language. The richness of the mathematics in a lesson was rated as low (1), medium (2), or high (3) on five sub-dimensions, which were then used to inform the overall rating for richness of mathematics. The five sub-dimensions were:

- linking and connection (e.g., of ideas, procedures, representations).
- teacher explanations (e.g., of why a procedure works, why a solution makes sense).
- multiple procedures or solution methods (for a single problem or a problem type).
- developing mathematical generalisations (e.g., examining cases and noting a pattern).
- mathematical language (fluent and explicit use of mathematical terms).

Comparing JUMP and IMPACT ratings, the sub-dimension on which they differed most notably was teacher explanations (Table 4.18). JUMP classes initially received a mean rating of 2.5 (i.e., medium to high quality) for teacher explanations, which rose to 2.8 for the second observations. Ratings for teacher explanations in IMPACT lessons also increased, but from an average of 1.8 (just below medium quality) to 2.3 (a little above medium).

Table 4.18: Observers' mean ratings of recorded lessons on the Richness of the Mathematics dimension, on a scale of 1 (Low) to 3 (High)

	JUMP (N=14 ratings, 7 classes)		IMPACT (N=16 ratings, 8 classes)	
	Time 1	Time 2	Time 1	Time 2
Teacher Explanations	2.5	2.8	1.8	2.3
Mathematical Language	1.8	2.3	1.6	2.4
Linking and Connection	1.7	2.0	1.6	2.0
Multiple Procedures or Solution Methods	1.3	1.5	1.5	1.4
Developing Mathematical Generalisations	1.3	1.7	1.3	1.5
Overall Richness of the Mathematics	1.6	2.1	1.4	2.0

Ratings for teacher use of mathematical language in JUMP lessons increased from 1.8 to 2.3, but from 1.6 to 2.4 in IMPACT. Across both programmes, ratings were lowest for the

use of multiple procedures and solution methods and for developing mathematical generalisations. The overall richness of the mathematics was estimated at a mean of 1.6 in JUMP and 1.4 in IMPACT (i.e., low to medium for both groups) for the first observations, but rose to 2.1 and 2.0, respectively, for the second observations.

Working with pupils and mathematics

In assessing how appropriately teachers responded to pupils' mathematical errors and productions, the two sub-dimensions rated were:

- remediation of pupil errors and difficulties (at procedural and conceptual levels).
- responding to pupil mathematical productions in instruction (e.g., identifying the mathematical relevance of pupil questions, using pupil ideas to build instruction).

As with the previous dimension, SME ratings suggested an improvement in quality between the first and second observations. The improvement was slightly more pronounced in IMPACT ratings, especially in how teachers responded to pupil mathematical productions – rising from 1.6 (low to medium) to a mean rating of 2.3 (medium to high) by the second observations (Table 4.19). On the overall scale for working with pupils and mathematics, the mean rating for JUMP classes was 2.1 (medium quality) on both occasions. For IMPACT, it was 1.7 (a little below medium quality), which rose to 2.0 (medium) for the second observations.

Table 4.19: Observers' mean ratings of recorded lessons on the Working with Pupils and Mathematics dimension, on a scale of 1 (Low) to 3 (High)

	JUMP (N=14 ratings, 7 classes)		IMPACT (N=16 ratings, 8 classes)	
	Time 1	Time 2	Time 1	Time 2
Remediation of Pupil Errors and Difficulties	1.8	2.0	1.4	1.8
Responding to Pupil Mathematical Productions	2.1	2.4	1.6	2.3
Overall Working with Pupils and Mathematics	2.1	2.1	1.7	2.0

Errors and imprecision

For the dimension of errors and imprecision, a low rating (1) was a positive rating (i.e., low level of teacher error), while a high rating of 3 was a negative rating. There were three sub-dimensions to the error and imprecision rating:

- major mathematical errors (e.g., solving problems incorrectly, omitting a key condition in a definition).
- imprecision in language or notation (e.g., errors in mathematical symbols or language).
- lack of clarity in presentation of mathematical content (e.g., mathematical point is muddled).

In the first set of observations, all teachers were rated as displaying low levels of major mathematical errors (i.e., a positive rating) (Table 4.20). Ratings were also almost uniformly positive for imprecision in mathematical language or notation, but slightly less positive for IMPACT teachers where clarity of presentation was concerned. Six (of the eight) IMPACT teachers were rated by at least one SME as having medium levels of clarity in presentation. Overall, though, the SMEs' ratings for errors and imprecision indicate low levels of errors evident during the first set of observations.

In the second set of observations, JUMP teacher ratings indicate a slight increase in errors and imprecision. For example, two JUMP teachers were rated by both SMEs as displaying a medium level of mathematical errors during the lesson (compared to no teachers during the first observations). The mean rating for lack of clarity also increased for JUMP lessons, while decreasing slightly for IMPACT. The overall errors and imprecision rating for JUMP lessons for the second observations was 1.2, compared to 1.1 for IMPACT lessons.

Table 4.20: Observers' mean ratings of recorded lessons on the Errors and Imprecision dimension, on a scale of 1 (Low level of error) to 3 (High level of error)

	JUMP (N=14 ratings, 7 classes)		IMPACT (N=16 ratings, 8 classes)	
	Time 1	Time 2	Time 1	Time 2
Major Mathematical Errors	1.0	1.3	1.0	1.1
Imprecision in Mathematical Language/Notation	1.1	1.3	1.1	1.2
Lack of Clarity in Presentation of Content	1.1	1.4	1.4	1.2
Overall Errors and Imprecision	1.1	1.2	1.1	1.1

Pupil participation in meaning-making and reasoning

Pupils were considered to participate in meaning-making and reasoning when they provided explanations, generated questions or arguments, and demonstrated engagement at a high cognitive level. Three sub-dimensions were rated, to contribute to the overall dimension:

- pupils provide explanations (may be pupil-initiated or teacher-initiated).
- pupil mathematical questioning and reasoning (e.g., pupils ask questions requiring explanations, make conjectures, or reason out conclusions).
- enacted task cognitive activation (i.e., whether pupils engage with tasks using a low, mixed or high level of thinking skills).

In the first observations, JUMP lessons were rated slightly lower than IMPACT lessons on pupils providing explanations (1.5 versus 1.8) (Table 4.21). By the second observations, this difference was reversed, with JUMP lessons averaging a rating of 1.9 (close to medium) for pupils providing explanations, compared to a rating of 1.7 for IMPACT lessons. Both groups received initial mean ratings of 1.4 for pupil mathematical questioning and reasoning (low to medium levels), and ratings for both rose slightly for the second observations.

Table 4.21: Observers' mean ratings of recorded lessons on the Pupil Participation in Meaning-Making and Reasoning dimension, on a scale of 1 (Low) to 3 (High)

	JUMP (N=14 ratings, 7 classes)		IMPACT (N=16 ratings, 8 classes)	
	Time 1	Time 2	Time 1	Time 2
Pupils Provide Explanations	1.5	1.9	1.8	1.7
Pupil Mathematical Questioning and Reasoning	1.4	1.6	1.4	1.7
Enacted Task Cognitive Activation	1.9	2.0	1.6	2.1
Overall Pupil Participation in Meaning-Making/Reasoning	1.6	1.9	1.7	2.2

JUMP lessons were initially rated slightly higher than IMPACT lessons for enacted task cognitive activation (1.9 versus 1.6). For the second observation, JUMP lesson ratings increased marginally, whereas IMPACT lessons rose by half a point (to just above medium). Initially, the ratings for overall pupil participation in meaning-making and reasoning were similar (1.6 for JUMP and 1.7 for IMPACT). However, in the second observations, IMPACT lessons averaged 2.2 (a little above a medium rating) whereas JUMP averaged 1.9 (or just below medium).

Global ratings of MQI and MKT

For each lesson viewed, each SME gave an overall estimate of mathematical quality of instruction as either low, medium or high. They also took a "lesson-based guess" at whether a teacher's mathematical knowledge for teaching (MKT) was low, medium or high. For both programmes, the overall average MQI rating for teachers increased from the first to the second observation. MKT estimates remained static for JUMP, but the SMEs rated IMPACT teachers' mathematical knowledge for teaching higher in the second observation than they did in the first (Table 4.22). For the second observations, quality of instruction was close to *medium* for both programmes. Teacher MKT was estimated as a little below medium (mean of 1.7) for JUMP teachers and medium (mean of 2.0) for IMPACT teachers.

Table 4.22: Observers' ratings of recorded lessons for Mathematical Quality of Instruction and estimated Mathematical Knowledge for Teaching scores, on a scale of 1 (Low) to 3 (High)

	JUMP (N=14 ratings, 7 classes)		IMPACT (N=16 ratings, 8 classes)	
	Time 1	Time 2	Time 1	Time 2
MQI	1.7	1.9	1.5	2.1
MKT	1.7	1.7	1.6	2.0

General comments on lessons

As noted earlier, a lesson report was completed by an SME for each recorded class. The lesson reports, in conjunction with informal conversations with the SMEs and those who conducted the live observations, provided supplementary qualitative data on how lessons were enacted in practice.

The SMEs noted that review-style lessons were common in both programmes and both sets of observations, but particularly in the first set. However, few lessons in either group ended with overall recaps of lesson content. A partial explanation may be that some teachers reported being nervous during the first recorded observations, but being less so during the second (also evident in higher MQI scores for the second observation). Teaching new material might be seen as more stressful than reviewing familiar material, suggesting that the emphasis on review was related more to nerves than to adherence to a review-heavy teaching approach.

The lesson reports also raised concerns that teachers sometimes adhered closely to the letter, but not the spirit, of their assigned programme. In JUMP, some lessons featured considerable repetition and practice, but lacked reference to the larger mathematical ideas behind the repeated steps. In a small number of cases, the teacher was so intent on the "step-by-step" approach that they rejected suggestions from pupils who had moved ahead or come up with an alternative (correct) solution method. One SME noted that the high level of decomposition of concepts and skills that *should* characterise JUMP lessons might be considered a form of differentiation to be used with weaker pupils. However, the fact that relatively little differentiation was evident was an indicator of low adherence to some key JUMP principles.

In IMPACT, some lessons drew heavily on concrete materials, but did not link the concrete practice with abstract mathematical concepts. In others, the social constructivist approach to analysing a problem and generating a shared solution seemed poorly understood. While the IMPACT programme promotes the idea that the teacher should not be the sole validator of knowledge and that all pupil responses should be valued, a few teachers' interpretation of this was to treat all pupil responses as equally *correct*. Several lessons featured pupils talking at length about problems, but the construction of a solution was not a strong feature of the discussion.

In both groups, almost half the lessons started and/or ended with a session of mental mathematics practice, often unrelated to the main lesson topic. In JUMP classes, this was typically achieved through fast-paced exercises and games using an interactive whiteboard. In IMPACT classes, it sometimes involved activities specifically recommended by the programme, e.g., the "counting choir" and "sound of a number" (counting can). The SMEs noted that this approach has been endorsed by many within the Inspectorate and that teachers may be trying to use mental mathematics as a means of improving number understanding and skill.

The lesson reports also showed that superficially similar teaching strategies could have quite different quality ratings. For instance, "station teaching" was used in several IMPACT classes in both sets of observations.[3] However, the SMEs viewed it as having less positive outcomes when used throughout an entire lesson than when used in between sessions of teacher-led introduction and recapping. They also felt that the amount of time allocated per station should (but did not always) take account of the task complexity and the amount of time needed for set up or explanation of the task. While station activities were usually drawn from the IMPACT manual, they were sometimes presented to pupils without context – a further instance of adhering to the letter of the programme, but not to its spirit.

Further, pupils in a few lessons seemed overly familiar with the station activities, suggesting that these observed lessons might have been repeated lessons. Indeed, in most cases it would have been impossible to run station-based lessons without an initial introduction, as pupils would not know what to do at each station. Some of those who conducted the live observations felt that at least some aspects of a small number of lessons had been rehearsed. Supporting evidence for this is found in some pupils' comments during interviews. For instance, when asked whether the lesson was typical, one pupil remarked that it was "*the same as most classes since two weeks ago*".

In a few other cases, observers inferred from conversation with teachers and pupils that the lessons they watched were atypical in their adherence to the assigned programmes. Some teachers commented that while they did not usually stick rigidly to their programme, they had made a particular effort for the observation. As will be outlined in Chapter 5, some pupils noted that the amount of games and activities used in observed lessons was unusual.

On balance, it may be that the observed lessons (particularly the first set) did not all represent typical mathematics lessons in the classes observed. A small number may have repeated certain elements from previous lessons, while others may have contained "*more props than normal*" (as one pupil succinctly put it). In a separate but related vein, the possibility that there could have been a similar approach to the end-of-year testing cannot be excluded. In two classes, observers expressed a concern that pupils might have been schooled on the DPMT.

[3] Although the IMPACT programme suggests activities for small group work, the manuals do not explicitly mention station teaching.

Summary

The observations were intended to provide measures of teachers' adherence to assigned programmes, and to examine the quality of instruction experienced by pupils in the two groups. Observer ratings indicated moderately strong adherence to assigned programmes. Most teachers in each group demonstrated *some* adherence to the relevant lesson plan or programme principles, but few demonstrated very close adherence, and a few appeared not to be basing their lesson on the assigned programme principles. Data from the observation schedule suggest that there was higher adherence to certain aspects of the programmes. In particular, the emphasis on teacher-led instruction and solo work was, as would be expected, more prominent in JUMP than IMPACT lessons, and group work and pupil discussion were more prominent in IMPACT than JUMP. However, prominence was relative to the other programme, and in most cases, not markedly high in a broader sense. For example, JUMP classes had more solo work than IMPACT classes, but not enough solo work to be considered outside the normal range of classroom practices.

Other aspects of programmes were less frequently observed than might be expected, particularly during the observations conducted later in the school year (e.g., JUMP bonus questions, use of memorisation and repeat procedures, use of workbooks). This, in conjunction with already noted concerns as to the representativeness of a few of the first observations, suggests that aspects of programmes might have been patchy in some classes. At a more qualitative level, the SMEs commented that in the recorded observations, there was relatively little evidence of the IMPACT programme's emphasis on a social constructivist approach or JUMP's intended reliance on frequent use of summative assessment.

Ratings of quality of instruction were similar across the two groups. Pupil engagement and *time on task* were high overall, while pupil understanding was slightly lower. Generally, quality ratings improved between the first and second set of recorded observations, perhaps due to teachers becoming more comfortable in front of a camera. JUMP and IMPACT lessons received similar ratings on most MQI dimensions. A notable exception was the quality of teacher explanations, which was higher on both occasions in JUMP lessons. Informal comments from SMEs and observers suggested that some observed lessons, particularly those in the first set, may not have always been an accurate reflection of a typical lesson.