GROUP WORK IN ELEMENTARY SCIENCE: ORGANISATIONAL PRINCIPLES FOR CLASSROOM TEACHING

Christine Howe*, Andy Tolmie*, Allen Thurston**, Keith Topping**, Donald Christie*, Kay Livingston*, Emma Jessiman** and Caroline Donaldson**

*University of Strathclyde, UK

**University of Dundee, UK

Address for Correspondence

Professor Christine Howe, Department of Psychology, University of Strathclyde, 40 George Street, Glasgow G1 1QE, UK. Fax: 0141 548 4001. Email: c.j.howe@strath.ac.uk

Abstract

Group work has been promoted in many countries as a key component of elementary science. However, little guidance is given as to how group work should be organised, and because previous research has seldom been conducted in authentic classrooms, its message is merely indicative. A study is reported, which attempts to address these limitations. Twenty-four classes of ten- to twelve-year-old pupils engaged in programmes of teaching on evaporation and condensation, and force and motion. Both programmes were delivered by classroom teachers, and made extensive use of group work. Pupil understanding progressed from pre-tests prior to the programmes to post-tests afterwards, and results suggest that group work played a critical role. Organisational principles are extrapolated from the findings, which could be readily adopted in classrooms.

1. Introduction

Over the past twenty years, group work between pupils has been promoted in many countries as a key component of elementary science. The relevance of group work to science education is a recurring theme in contemporary guides for practitioners (e.g. Harlen & Qualter, 2004; Sharan, 1999; Ward, Roden, Hewlett, & Foreman, 2005). It is emphasised throughout a recent issue of the professional journal *Primary Science Review*, which addresses ‘questions and dialogue’ (Association for Science Education, 2004). Within the United Kingdom, it has even reached the level of national policy. A crucial version of the science curriculum for England and Wales (Department of Education and Science, 1989) stipulates that ‘pupils should describe and communicate their observations, ideally through talking in groups’ (p.3). *Environmental Studies: Science* (Learning and Teaching Scotland, 2000), which guides teaching in Scotland across the first nine years of schooling, also recommends group work, stating that it is ‘particularly suitable for encouraging pupils to talk about and share ideas’ (p.20).

Yet seldom, in either the practitioner publications or the policy documents, is detailed guidance given about how science group work should be organised. At best, suggestions are made about the broad types of activity that are amenable to group work, e.g. practical investigation (Sharan, 1999). However, the use of group work for practical investigation is already well established in elementary science (Lemke, 1990). Since current group work practices (including for science) have been criticised for being undemanding (Baines, Blatchford, & Kutnick, 2003; Bennett, Desforges, Cockburn, & Wilkinson, 1984; Galton, Hargreaves, Comber, Wall, & Pell, 1999; Galton, Simon, & Croll, 1980), advocacy of practical work seems unlikely to be sufficient. Advice for elementary science that goes beyond the level of broad activity would appear to be necessary. The over-arching aim of the research to be reported here is to contribute to such advice, via the specification of relatively fine-grained principles for designing effective group work.

1.1. Organisational principles

Theoretical analyses of classroom group work can usually be traced back to Dewey's (e.g. 1916) contention that pupils should be encouraged to operate as members of communities, actively pursuing interests in cooperation with others. For instance, inspired by Dewey, Piaget (e.g. 1932) saw cooperation as providing the social context where pupils would be motivated to coordinate existing ideas with alternatives. Such coordination was, for Piaget, the precondition for development. Piaget's emphasis upon cooperation as a means to coordination led him to argue also that group work between pupils should be more productive than interaction with adults. Ideas proposed by adults would, according to Piaget, most likely dominate those held by pupils rather than be coordinated with them. Subsequently, researchers in the 'cooperative learning' tradition (e.g. Johnson & Johnson, 1992; Slavin, 1995) have revisited Dewey's emphasis upon the joint pursuit of interests. While accepting the need for exchange of ideas as highlighted by Piaget, they focus upon the opportunities provided by group work for pooling resources towards individual (but interdependent) goals.

Whatever their differences, the background theories all highlight interaction between pupils in contexts of mutuality and equality, and this emphasis has provided the framework for much subsequent research. In general, results have concurred with the framework, while fleshing it out. For instance, research has covered *structural features* like group size and seating arrangements. Consistent with the framework, it recommends groups of no more than four or five pupils (e.g. Baines et al., 2003; Lou, Abrami, Spence, Poulsen, Chambers, & d'Apollonia, 1996), and equal opportunities for eye contact (e.g. Jaques, 2000). A second strand in previous research relates to the *role of teachers*, looking in particular at desirable forms of intervention and optimal principles of task design. The basic message is that intervention should stress monitoring and guidance rather than control (e.g. Blatchford, Baines, Rubie-Davies, Bassett, & Chowne, 2006; Cohen, Lotan, & Leechor, 1989), and tasks should be designed to be open-ended, challenging and inherently cooperative (e.g. Cohen, 1994; Slavin, 1995). Again, this concurs with the background framework.

Research has also addressed productive *forms of pupil interaction*, with work conducted in a range of contexts. Studies have addressed mathematics (Damon & Phelps, 1989; Webb, 1989), the humanities (Miell & MacDonald, 2000; Morgan, Hargreaves, & Joiner, 2000), and the social sciences (Shachar & Sharan, 1994). Of particular relevance here, they have also covered elementary concepts in science, e.g. biological transmission (Williams & Binnie, 2002; Williams & Tolmie, 2000), heat transfer (Howe, Tolmie, Greer, & Mackenzie, 1995; Howe & Tolmie, 2003), object flotation (Howe, Rodgers, & Tolmie, 1990; Howe, McWilliam, & Cross, 2005), propelled motion (Hennessy, Twigger, Driver, O'Shea, O'Malley, Byard, Draper, Hartley, Mohamed, & Scanlon, 1995; Howe, Tolmie, & Rodgers, 1992) and shadow formation (Forman & Carr, 1992; Howe, Tolmie, Duchak-Tanner, & Rattray, 2000). Regardless of context, optimal interaction seems to involve pupils proposing ideas and explaining their reasoning to their peers, perhaps after disagreement. This of course is highly consistent with the Piagetian emphasis upon the exchange of ideas. Going beyond Piaget (although not incompatible with his approach), it also appears helpful, but not essential, for pupils to refer back to earlier task material, or when the task is challenging, to achieve group consensus over crucial points.

Given the extent of previous research, it might appear that specifying principles for effective group work in elementary science is simply a matter of disseminating what is already known. However, this is not the case. In the first place, the work on structural features and the role of teachers has, with very few exceptions (e.g. Sharan, 1999), focused upon disciplines other than science. Yet science classrooms have characteristics, for instance extensive use of equipment, which mean that principles established elsewhere will not necessarily apply. As regards the third area covered above, i.e. pupil interaction, research has, as noted, included elementary science. However, again with a small number of exceptions (e.g. Barnes & Todd, 1977; Herrenkohl, Palincsar, DeWater, & Kawasaki, 1999), it has not been classroom-based. It has usually followed an experimental methodology, and although this has resulted in large samples of pupils, controlled assessments of impact, and evidence of sustainable benefits, it has also involved short-term (often one-off) interventions, researcher- rather than teacher-delivery, artificially constructed groups, and unfamiliar settings. The dangers of extrapolating from experimental findings have been highlighted by Engestrom (e.g. 2004) and Kumpulainen, Kangassalo and Vasama (2005), who warn in particular about insensitivity to local conditions and/or contextualised meanings. There is, as a consequence, a need for research that checks whether principles of effective group work *suggested* in experimental contexts have relevance in standard classrooms. The research that follows was intended to address this need.

1.2. SPRinG KS2

The research to follow builds on SPRinG KS2 (Social Pedagogic Research into Group Work Key Stage Two), a recent project (Blatchford et al., 2006) that, in contrast to work reviewed so far: a) did take place in authentic classrooms, b) was concerned with elementary science, *and* c) did cover structural features, teacher role, and pupil interaction. The project was conducted in the South of England with fourth-, fifth- or sixth-year pupils, Key Stage Two covering the second half of what, in the United Kingdom, is called ‘primary’ schooling. The project was multi-faceted, but crucial for present purposes are the following. First, it started with activities for developing generic group skills, which were introduced in class by teachers working from resources that the researchers provided. The skills included listening, questioning, helping, giving explanations and reaching agreement. Second, subsequent to skills training, the pupils went through two programmes of science teaching, one addressing evaporation and condensation, and the other addressing force and motion. Each programme covered key concepts, and required pupils to design investigations. For instance, the force and motion programme covered the angle, smoothness and height of slopes, and the weight and streamlining of cars as influences on motion, and introduced the concepts of gravity, friction and air resistance.

Although the science programmes employed whole-class discussion and teacher demonstration, they made extensive use of group work. The group tasks incorporated features shown in earlier experimental studies (primarily Howe & Tolmie, 2003; Howe et al., 1995, 2000; Tolmie, Howe, Mackenzie, & Greer, 1993) to maximise the chances of pupils proposing ideas, disagreeing, explaining their reasoning, referring back and reaching consensus. In other words, the tasks were designed to support the forms of pupil interaction that the studies (as well as other work reviewed above) had found to be beneficial. The programmes were implemented by teachers using researcher-supplied resources (which had themselves been developed in consultation with teachers), and in each case involved two to three hours of teaching spread over several weeks. Pupil understanding of evaporation and condensation and force and motion was tested before and after the programmes, and progress significantly exceeded that made by control pupils who received teaching in the two topic areas, but did not participate in the group skills training or the SPRinG science programmes.

The SPRinG KS2 project is an important step towards clarifying how group work should be organised in elementary science, yet uncertainties remain. First, there is no guarantee, from the patterns of pre- to post-test change alone, that the group work component of the science programmes, and in particular interaction of the form highlighted by past research, was a critical factor in promoting growth. The benefits may have resulted from other aspects of the programmes, e.g. the whole-class discussions or teacher demonstrations. They may also have stemmed from the group skills training, and not the science programmes. It is possible, for instance, that the training boosted pupil motivation and interest, guaranteeing a positive response to science teaching no matter how it was presented. Observational data were collected while the programmes were being implemented, and these data may clarify which aspects were beneficial. Nevertheless, designed for different purposes, the observational categories do not correspond precisely with notions like proposing, disagreeing, explaining, referring back and reaching consensus, and therefore results may not be conclusive.

Second, even if the group work component was important, there is no guarantee that the benefits will generalise across classroom contexts. As noted earlier, the background theories emphasise equality and mutuality, implying that group work should involve pupils of roughly equal status. Supporting this is evidence that group members are most likely to share ideas and explain their reasoning when status is similar (Kruger & Tomasello, 1986; Mugny & Doise, 1978). With asymmetries, higher status pupils will often dominate (Bachmann & Grossen, 2004; Ellis & Rogoff, 1982; Miller & Brownell, 1975). SPRinG's science programmes focused on single-age classes, and age is an important (although not unique) predictor of pupil status (Rogoff, 1990). Some schools in the United Kingdom deploy mixed-age composite classes, as do schools elsewhere in Europe and North America. Composites (often with very wide age ranges) are the norm in developing countries. Questions must therefore be raised about the viability of the SPRinG group tasks in such contexts.

1.3. Present research

In view of the above, the research that is reported here deployed materials modelled on the SPRinG science programmes to address two questions. The first question was whether use of the materials is associated with knowledge gains in composite as well as single-age classes. The research was conducted in Scotland, where the reasons for composite classes differ between rural and urban schools. In rural schools, composites often occur because the numbers of pupils in any given age band are below those needed to comply with guidelines on teacher-pupil ratios. In urban schools, they usually occur when the numbers are too high, and compliance can only be achieved via single-age classes at each of two age levels plus mixed-age composites. In any event, the smaller staff complement in rural schools may mean greater variation in teachers who are confident and able to tackle science. Acknowledging such issues, the research counter-balanced the single-age vs. composite contrast, with a rural vs. urban contrast.

The second question addressed in the research was whether knowledge gains after the science programmes are dependent upon group work that displays the structural features, teacher contribution and pupil interaction that studies discussed earlier suggest may be beneficial. Structural features and teacher contribution were examined using rating scales, while pupil interaction was explored using rating scales together with classroom observations that looked directly at proposing, disagreeing, explaining, referring back and reaching consensus. To allow the contribution of group work to be pinpointed as precisely as possible, observations were made during teaching sessions that were mainly group-based *and* during sessions that emphasised whole-class activities directed by the teacher. To double-check that the group-based and whole-class sessions contrasted as expected, records were also made of the social context in which pupils were located, e.g. working alone, with the teacher, or in a group.

2. Method

2.1. Design

The research was part of a larger project, which took place between August 2003 and June 2004, and replicated all aspects of the SPRinG KS2 project. Thus, the research included extensive training in generic group skills between October and December, and summative assessments of school performance, self-appraisal, peer relations and attitudes to schooling during October and June (Christie, Tolmie, Howe, Topping, Thurston, Livingston, Jessiman, & Donaldson, 2005; Thurston, Howe, Christie, Topping, Tolmie, Livingston, Jessiman, & Donaldson, 2005). Of relevance here are the two science programmes that were implemented between February and April, the observations and ratings of group work that were made during implementation, and the pre- and post-tests that were used to assess benefits.

The pre- and post-test scores of pupils who participated in the science programmes (intervention pupils) were compared with the scores of control pupils who were not only excluded from the programmes but also, in contrast to the SPRinG KS2 pupils, not taught the topics by other means. A non-instructed control group was appropriate in the present context, because its function was purely to clarify whether the programmes were effective in boosting scores. Given evidence of effectiveness, the first research question was addressed by comparing the pre- and post-test scores of the intervention pupils as a function of type of class, i.e. single-age or composite in rural or urban locations. The second question was addressed via the extent to which post-test scores could be predicted by observations and ratings. Thus, pre- and post-test scores were the dependent variables, and intervention or control, type of class, and observations and ratings were the independent variables.

2.2. Sample

Details of the project were distributed to 221 primary schools in eight local authority regions in central Scotland. Four regions were located in the eastern part of the country, and four were located in the western part. Twenty-four schools were selected to participate in the intervention from the 85 expressing interest, with selection considering region, rural vs. urban, and single-age vs. composite but otherwise random. In particular, at least one school was chosen from each region, with twelve of the selected schools located in the eastern regions, and twelve in the western. Working from the 2001 Census (General Register Office for Scotland, 2004), the schools' postcodes were used to determine whether their local population densities were more (urban) or less (rural) than 10,000 people. The 10,000 cut-off corresponds with Government categorisations. Half of the schools in both the eastern and western regions were in urban locations, and half were in rural. Government records were used to determine the percentage of pupils in each school in receipt of free school meals, as a straightforward index of socio-economic status.

One class comprising fifth-, sixth- and/or seventh-year pupils (aged ten to twelve years) was selected from each of the 24 schools, to give three single-age classes and three composite classes in each of the eastern urban, eastern rural, western urban and western rural locations. The distinction between single-age and composite classes was made on the basis of whether or not the pupils in each class had started primary school in the same school year. The number of fifth-, sixth- and seventh-year pupils was balanced across the single-age and composite classes, and the urban and rural locations. Intervention sample details are summarised in Table 1, together with information about three control classes. Control schools proved difficult to recruit due to understandable reluctance to be excluded from the intervention, and therefore the control classes are single-age only **[Insert Table 1 about here]**.

2.3. Materials

All materials were adapted from those used in the SPRinG KS2 project, with modifications driven by Scottish curricular demands, the slightly older age group, and the specific aims of the present research.

2.3.1. Science programmes

The research involved two programmes of science teaching, one on evaporation and condensation, and the other on force and motion. Both programmes provided comprehensive coverage of the relevant curriculum, while being consistent with the time periods that would, within the United Kingdom, normally be devoted to single topics in science. The programmes were detailed in notes for teachers, with group activities outlined in booklets for pupils. The latter incorporated features that had been found in the experimental studies of Howe and colleagues (Howe & Tolmie, 2003; Howe et al., 1995, 2000; Tolmie et al., 1993) to maximise the likelihood of proposing, disagreeing, explaining, referring back and reaching consensus. These features amounted to instructions that pupils share ideas about how events should be explained, discuss their ideas and reach consensus, and record the agreed explanation in writing.

The evaporation and condensation programme was in three parts, with the notes indicating that the shortest part should take about 70 minutes and the longest part about 95 minutes. A few days before Part 1, which focused on evaporation, a tank of water was to be set up in the classroom, and the water depth to be periodically recorded. Part 1 was to begin by highlighting the water's disappearance, together with the (demonstrated) disappearance of wet handprints from kitchen roll. Working in groups, pupils were then to discuss, agree and record their views about what happens to the water in towels as they dry on washing lines. The term 'evaporation' was to be introduced at this point. Part 1 was to continue with perfume poured into a dish, and the observation that the pupils who were closest smelled the scent first. After discussing why this happened, groups were to revisit their theories about drying towels. Part 2 was to start with each group identifying their best idea about the towels, and subjecting the idea to empirical test. Controlled investigations were to be planned and implemented, and results were to be recorded and interpreted. The theme of Part 3 was condensation, with the concept (and term) to be introduced via the water that appears when steam from a boiling kettle hits a metal dish. Working in groups, pupils were to discuss, agree and record their views about where the water comes from. They were then to be introduced to the distinction between solids, liquids and gases, via playground exercises that simulate molecular structure (e.g. crowding together with tightly linked arms for solids). In groups, pupils were to use the distinction to interpret evaporation and condensation from the kettle. Finally, groups were to note (and interpret) that when mirrors are placed above hot and cold water, mist forms on the hot water mirror alone.

The force and motion programme was in seven parts, with the shortest part expected to last about ten minutes and the longest part about 105 minutes. In Part 1, pupils were to observe a toy car rolling down a slope, and in groups, to discuss, agree and record how (via changes to the slope's height or surface, or the car's shape), they could make the car roll as far as possible off the slope. In Part 2, groups were to find as many ways as possible of moving a toy car across a table, and discuss, agree and record whether the car could be moved without a push or pull. The notion that forces are kinds of pushes or pulls was to be introduced. Part 3 was to focus on the force of gravity, with observations of dropped objects and discussion of how gravity pulls the toy car down the slope. Consideration was to be given to forces that prevent falling by opposing gravity, e.g. the force from a hand on a held ball. Part 4 was similar to Part 1, except that now groups were to discuss, agree and record how they could make a rolling car travel as short a distance as possible. The concept of 'friction' could be introduced, but whether it was or not, friction was to be the focus of Part 5. Groups were to discuss friction, and plan fair tests of how slope surface impacts upon rolling. Tests were to be carried out, and results interpreted. The theme of Part 6 was 'air resistance', with the notion to be introduced via demonstrations of how two identical cotton wool balls fall with the same speed if both are compressed, but at different speeds if one is fluffed out. Working in groups, pupils were to agree and write down explanations of why this happens. The role of air resistance was to be reinforced via discussions of the falling speeds of small, medium and large parachutes (made from bin liners, string and paper-clip weights). Part 6 was to end with planning, implementing and interpreting fair tests of how air resistance affects rolling down slopes (via toy cars with or without pieces of card attached to their fronts). The programme was to conclude in Part 7 with 'the great down the slope car race', where groups were to prepare slopes and cars so that their own car would travel as far as possible, and the teacher's car would travel as little as possible.

2.3.2. Pre- and post-tests

Both programmes were associated with pre- and post-tests, presented in booklet form for written response, and covering all aspects of the programmes. There were 16 questions in the pre- and post-tests for the evaporation and condensation programme (15 multiple-choice and one open-ended). Sometimes respondents were instructed to select one multiple-choice option only; sometimes they were told that several might be correct. Each pre-test question had a post-test equivalent, with the tests differing only in their problem content, and the order in which multiple-choice options were presented. Seven questions focused upon evaporation, e.g. a) where has the water gone from towels (pre-test) or clothes (post-test) drying on washing lines (five options, e.g. 'soaked in', 'gone into the air'); b) what happened to the water (five options, e.g. 'became a gas', 'disappeared'); c) what happened to the molecules in the water (seven options, e.g. 'far apart and moving quickly', 'close together and moving slowly’); d) what is the process called (six options, e.g. 'condensation', 'evaporation'). Four questions addressed tests to explore whether hanging clothes outside (pre-test) or smoothing clothes out (post-test) helps them dry. They covered: a) manipulating key variables (four options, e.g. 'hang one (of two) piece outside', 'fold one piece up'); b) holding other variables constant (seven options, e.g. 'use different kinds of material', 'use the same amount of water to wet them'); c) necessary apparatus (four options, e.g. 'clock', 'ruler'); d) predicted outcomes (open-ended). Five questions asked about condensation, e.g. a) why does water appear on the outside of bottles from the fridge (pre-test), or on the inside of windows in winter (post-test) (five options, e.g. 'water soaks through', 'cold makes water vapour turn into a liquid'); b) what happens to the molecules (five options, e.g. 'nothing', 'lose a lot of energy and move close together'); c) what is the process called (six options, e.g. 'disappearing', 'condensation').

There were 29 multiple-choice questions in the force and motion pre- and post-tests, and again the two tests were identical apart from the ordering of options and minor aspects of content. Fourteen questions addressed the properties of slopes and cars relevant to speed of rolling. The questions were associated with diagrams, each of which showed a pair of slopes and cars. The pairs varied along one dimension (e.g. high or middle slope), two dimensions (e.g. bumpy or smooth slope, medium or big push), or three dimensions (e.g. high or middle slope, bumpy or smooth slope, pointed- or flat-fronted car). The task with each diagram was to identify which car would roll furthest, and why this would happen (five options, e.g. 'the car is lighter', 'there is less air resistance'). Twelve questions focused on forces. Definitions were requested for gravity (five options, e.g. 'air pushing down', 'the pull of one object on another'), friction (five options, e.g. 'the rubbing of one surface against another', 'wind blowing against an object') and air resistance (five options, e.g. ‘the push of air against an object', 'wind pushing an object along'). Then each force was to be drawn on a diagram, which showed a car on a slope, and explanations were to be given of what the force does (two options, i.e. 'make the car move', 'slow the car down'), and how it works (four options, e.g. 'rubs back against the wheels', 'the air pushes the car down'). The tests ended with three questions where, due to combinations of slope and car characteristics, the car could be said to be moving quickly, moving slowly or not moving. Explanations were to be given (five or six options, e.g. 'the car is very heavy', 'there is less air resistance on the car').

2.3.3. Observations and ratings

Classroom observations were to be made *in situ* using a time-sampling methodology, with researchers employing grids to record pupil behaviour. The rows in the grids corresponded to the sampled periods, and the columns corresponded to the behavioural categories. Four categories related to the social context in which pupils were located: a) alone - working on their own; b) teacher - engaged with (i.e. talking or listening to) a teacher or classroom assistant; c) own group - engaged with another pupil in the same group or in close proximity; d) other group - engaged with a pupil in a different group. The remaining categories covered key aspects of dialogue: e) proposition - suggests an idea or course of action; f) disagreement - rejects another’s suggestion or explanation; g) explanation - gives a reason for a proposition; h) reference back - refers to a previous suggestion or explanation; i) resolution - adjusts to/agrees with another's previous statement; j) instruction - tells someone to say or do something; k) question - asks an open-ended question (or gives another form of prompt); l) uncodable - inaudible or not covered by the above categories. Videotapes previously recorded in elementary science classrooms were used to train two researchers in use of the categories, and reliability checks were conducted via the independent coding of 64 forty-second extracts from further videotapes. Inter-researcher agreement over the categories ranged from 82% to 100% (*M* = 92%). Prior to collecting the reported data, the researchers used the observation grids, on two occasions each, in all intervention classrooms.

Ratings of the group activity that took place across full teaching sessions were to be obtained using an instrument referred to as ‘S-TOP’ (SPRinG-Teaching Observation Protocol). S-TOP comprised 31 scales, with each scale requesting that a statement, e.g. ‘The size of groups maximised pupil-pupil interaction’, ‘The teacher spent time monitoring the groups’, be rated 1 if *not true* of the session, 2 if *partly true,* and 3 if *very true.* All statements were worded such that ratings of 3 indicated group work that was consistent with recommendations from the background literature. Four 'learning context' scales related to structural features, covering the appropriateness of seating arrangements, size of groups, number of groups, and pupil organisational behaviour. Seven 'activities and tasks' scales addressed the extent to which the materials encouraged group work, allowed pupils to organise their work, supported explanatory discourse, allowed the achievement of consensus, were open-ended, were appropriately structured, and were understood by pupils. Nine 'role of adults' scales covered teacher encouragement to appropriate time management, briefing on working in groups, modelling of good interaction skills, encouragement to use of group skills, monitoring, non-directive guidance, debriefing, encouragement to reflection during debriefing, and evaluation of group work during debriefing. Eleven 'group interaction' scales covered full pupil involvement, avoidance of sub-groups, extent of on-task talk, positive attitude, sharing and developing ideas, use of 'exploratory talk' (Mercer, 2000), constructive evaluation, achievement of consensus or compromise, productive discussion or conflict, good conversational skills, and appropriate role divisions. The S-TOP scales therefore covered group structure, teacher input *and* pupil interaction, while the observation grids focused upon pupil interaction alone.

2.4. Procedure

The science programmes and pre- and post-tests were introduced to the relevant classroom teachers during a continuing professional development session. The observation grids and S-TOP scales did not require lengthy description, since the teachers were familiar with them from records made during the group skills training that had taken place previously. The teacher notes and pupil booklets were distributed, and an overview was provided of each science programme in turn. Appropriate equipment was demonstrated, e.g. slopes and parachutes for the force and motion programme, but not supplied: the interest was in the viability of the programmes using materials routinely available in classrooms. The teachers were asked to begin with the evaporation and condensation programme, and take the force and motion programme second. They were requested to cover the programmes in full, keep to the intended sequence, and implement all group activities. However, they were also encouraged to adapt the programmes to the specific needs of their class, spending more or less time on any single activity as required. Although the designated parts of each programme could correspond to a single lesson, this was not essential: fewer or more lessons were equally acceptable. The teachers were asked to administer the pre-test at the start of each programme, and to administer the post-test two weeks after its conclusion. Control class teachers were introduced to the pre- and post-tests via individual sessions within their schools. Schedules for administration were agreed that ensured comparability with intervention classes. Inevitably, some intervention and control pupils were absent during pre- or post-testing, and these pupils were excluded from data analysis (see Table 1 for Ns on which analyses are based).

A researcher kept in touch with the intervention schools via telephone or email while the programmes were being implemented, and visited these schools for four observation sessions. Two sessions were scheduled to coincide with mainly whole-class teaching, one to cover evaporation and condensation and the other to cover force and motion. During the other two sessions, the emphasis was on group work, again with coverage of both science programmes. Prior to the first session, six pupils (three boys and three girls) had been randomly identified for observation from the class list. In composite classes, these target pupils included equal numbers from both age groups. The same six pupils were observed during all four sessions. Observations were based on 40-second windows, i.e. 12-seconds to prepare, 16-seconds to observe, and 12-seconds to record. In each session, eight successive windows were observed for one target pupil before moving to the second target pupil (and so on). Thus, each pupil was observed for exactly the same time as every other pupil within and across sessions. During group work, pupils were only observed when they were actually supposed to be conducting group activities (not during briefing or debriefing).

Observations were recorded via ticks on a grid, with separate grids used on each occasion for each target pupil. Multiple codes were used where appropriate, for all observations falling within the same window. For example if a pupil gave an instruction and then asked an open-ended question, both were recorded. For simplicity, each social context was recorded once only during a given window, no matter how many times the target pupil engaged with a specific partner. For instance, if the pupil started by talking with another pupil in the same group, then asked the teacher a question, and then returned to talking with the first pupil, this would be recorded as one ‘own group’ and one ‘teacher’. Dialogue variables, by contrast, were recorded every time they occurred, leading occasionally to repeated uses of codes within single windows. With the group work sessions only, the researcher used the S-TOP instrument after observing all six target pupils, to record her overall impressions of how the session had proceeded.

2.5. Scoring

As noted, there were 16 questions in the evaporation and condensation pre- and post-tests, and 29 questions in the force and motion pre- and post-tests. However, as intimated to the pupils, the best responses would, on occasion, have involved selecting more than one option. For instance, asked to use pieces of material to establish whether hanging outside speeds drying up, it would be good to use material of the same size *and* dampened with the same amount of water. Therefore, the best response to this question would tick both of these options. As a result, the maximum score achievable was 19 for the evaporation and condensation pre- and post-tests, and 37 for the force and motion pre- and post-tests. The pre- and post-test responses from each pupil were scored accordingly.

As regards the classroom observations, the observational grids produced pupil-level data, and S-TOP produced classroom-level data. Specifically, there were four grids for each of six target pupils in each intervention classroom, two grids relating to whole-class teaching (Class 1 and Class 2), and two relating to group work (Group 1 and Group 2). Each grid comprised observations across eight windows. As a result, coding involved counting the frequency with which each target pupil used each of the twelve categories within each grid. With S-TOP, there were ratings from each of Group 1 and Group 2, with the ratings addressing the learning context (4 scales, possible score range = 4 to 12), activities and tasks (7 scales, range = 7 to 21), role of adults (9 scales, range = 9 to 27) and group interaction (11 scales, range = 11 to 33). Taking Group 1 and Group 2 separately, scores were totalled across the scales for each dimension, with high scores corresponding to what the background literature defines as favourable ratings.

3. Results

The first aim of the research was to explore the effectiveness of the science programmes in composite classes as well as single-age classes, taking account of possible differences between rural and urban contexts. Addressing the aim was regarded as a two-stage process, involving comparisons between the pre- and post-test scores of first the intervention and control classes, and then the four types of intervention class. However, it was recognised that both comparisons might be confounded by differences in pupil age, gender, and socio-economic status. Therefore preliminary analyses were conducted to examine this possibility. There were no significant differences in the gender composition of the intervention and control classes, but the control pupils (M = 11.24 years) were significantly older (t (641)[1] = 5.76, $p < .001$) than the intervention pupils (M = 10.83 years). They were also of significantly lower socio-economic status (t (641) = 9.00, $p < .001$), as indexed by the percentage of free school meals (M for control = 28.43%; M for intervention = 17.58%). As a result, age and the socio-economic index were included as covariates in two-way, mixed-model ANOVAs, comparing pre- and post-test scores (within-subjects factor) in the intervention and control classes (between-subjects factor). There were no significant main effects (including for the covariates), and no significant interaction effects involving the covariates. However, there were statistically significant interactions between pre/post and intervention/control for both evaporation and condensation (F (1, 575) = 31.30, $p < .001$) and force and motion (F (1, 516) = 9.75, $p < .01$). As can be seen in Table 2, the reason for these results was consistent across the programmes: the mean scores of the intervention pupils improved from pre- to post-test, while the mean scores of the control pupils did not change. It can be concluded that the science programmes were both effective **[Insert Table 2 about here]**.

[1] Discrepancies in degrees of freedom across analyses stem from the fact that while preliminary analyses were conducted on all pupils registered in the participating classes, main analyses were, as noted, restricted to those pupils for whom pre- and post-test data were both available.

Preliminary analyses revealed no significant age or gender differences between the four intervention class types. However, a two-way (single-age vs. composite; urban vs. rural) ANOVA revealed that socio-economic status was significantly lower in the schools with composite classes (M = 20.94%) than the schools with single-age classes (M = 14.20%), and in the urban schools (M = 20.72%) than the rural schools (M = 13.69%). As a consequence, the socio-economic index was used as a covariate in the main analyses, which involved three-way (pre vs. post (within-subjects); single-age vs. composite (between-subjects); urban vs. rural (between-subjects)) mixed-model ANOVAs. The differences between pre- and post-test scores were significant for both evaporation and condensation (F (1, 509) = 63.31, p < .001) and force and motion (F (1, 460) = 43.10, p < .001). From Table 2, it is clear that the differences resulted from *gains* between pre- and post-test. There were no significant interactions between pre/post and any other variable (including the covariate), showing that the gains were independent of the classes’ status as single-age or composite and urban or rural. It can be inferred that the programmes were effective regardless of type of class.

The second aim was to explore whether the properties of group work recommended by previous research had any bearing on the positive outcomes. The observations made in the intervention classrooms addressed the properties of pupil interaction, while providing information about the social context within which interaction occurred. As Table 3 shows, the whole-class and small-group settings differed as intended in the interactive contexts they gave rise to. Group work, as indexed by the 'own group' category, was significantly more frequent during Groups 1 and 2 than during Classes 1 and 2, and although group work occurred during Classes 1 and 2, its frequency was dwarfed by the frequency of working alone. Group work was also associated with higher usage of the putatively helpful behaviours: Table 3 shows that propositions, disagreements, explanations, and references back all occurred more frequently in Groups 1 and 2 than in Classes 1 and 2. However, instruction and questions were also more frequent in Groups 1 and 2, and there was no expectation that these would prove helpful. A multivariate ANOVA revealed that use of the dialogue variables differed across the four observational settings (F (24, 1092) = 12.50, $p < .001$), and follow-up univariate ANOVAs summarised in Table 3 showed that significant differences occurred with six variables. Post hoc tests (Scheffé $p < .05$) on each of the six variables indicated that the frequencies in Groups 1 and 2 differed significantly from the frequencies in Classes 1 and 2, but there were no differences between the two Group settings or between the two Class settings.

[Insert Table 3 about here]

Having established that Groups 1 and 2 were associated with heightened frequencies of most dialogue variables, the key issue was whether use of these variables within Groups 1 and 2 predicted knowledge gains. As a preliminary, exploratory factor analysis was conducted on the frequencies of all dialogue variables, except uncodable, across Groups 1 and 2. After rotation, five factors emerged with eigenvalues greater than one, jointly accounting for 61.68% of the variance. However, only the first factor (eigenvalue = 2.97, 21.23% of the variance) was interpretable: it was defined by propositions in Group 1 (loading =

.82), explanations in Group 1 (loading = .73), propositions in Group 2 (loading = .80) and explanations in Group 2 (loading = .65). Accordingly, these four variables were combined into a single 'proposition/explanation' variable. Disagreement loaded weakly on the factor too, but the loadings (.45 for Group 1; .23 for Group 2) were insufficient to warrant combination. Thus, disagreement was treated separately, as were all other remaining variables. As Table 4 shows, only proposition/explanation was consistently associated with post-test score, when correlations were computed for the observed (i.e. target) pupils[2]. Furthermore, its value was unique to the group context. A multiple regression analysis was conducted to examine the extent to which post-test scores were predicted by the proposition/explanation variable when computed across Groups 1 and 2 and the same variable when computed across Classes 1 and 2. Proposition/explanation computed across Groups 1 and 2 predicted post-test score for both evaporation and condensation (β = .28, t = 3.10, p < .01), and force and motion (β = .29, t = 3.13, p < .01). However, it had no predictive value when computed across Classes 1 and 2 (for evaporation and condensation, β = .03, t = 0.36, *ns*; for force and motion, β = .12, t = 1.30, *ns*). Since propositions and explanations were identified as critical in previous research, recommended variables have therefore proved relevant in the present context. **[Insert Table 4 about here]**

[2] Correlations between dialogue variables and post-test scores are preferable over correlations between dialogue variables and pre- to post-test change (i.e. post-test less pre-test), even though *conceptually* the latter are of interest. This is because post-test scores are mathematically equivalent to pre- to post-test change after controlling for pre- test score (and therefore for any associations between pre-test and dialogue). In reality, analyses based on pre- to post-test change produced results that were similar to those shown in Table 4.

Unlike the observational data, which were restricted to pupil interaction, the S-TOP ratings covered interaction, group structure, and the role of teachers. From Table 5, it is clear the ratings were skewed towards the positive end of the scale in both Group 1 and Group 2, although the standard deviations attest to variation between classrooms for all but the learning context. The ratings were also highly correlated across Groups 1 and 2, suggesting that they could be combined for subsequent analyses. These analyses involved computing regression coefficients to examine whether the S-TOP dimensions predicted post-test scores for each of the science programmes. Because the S-TOP ratings were whole-class measures, the analyses addressed their ability to predict the mean post-test scores obtained by averaging across the pupils in each of the 24 intervention classrooms, rather than the scores achieved by the pupils taken separately. The results are shown in Table 5, where the scales associated with the role of adults can be seen to be strongly predictive. The more that the adults approximated a supportive, non-directive role, the more the pupils learned. The results from the other scales were in the anticipated direction, since all beta values were positive. However, except for the role of adults, no values were statistically significant. **[Insert Table 5 about here]**

4. Discussion

The research addressed two questions, in the hope that the answers would assist in specifying the principles around which elementary science group work should be organised. The first question was whether science programmes along the lines of those employed in the SPRinG KS2 project could be used successfully in composite classes as well as single-age classes, with urban and rural locations controlled for. Data summarised in Table 2 provide an unequivocally affirmative answer. Significant gains in understanding of both evaporation and condensation and force and motion were observed in the pupils who participated in the science programmes, while the control pupils made no progress whatsoever. Moreover, the gains in the participating pupils were constant across composite and single-age classes and urban and rural locations. Based on these results, it can be concluded that the SPRinG approach provides a robust framework for supporting elementary science, a robustness emphasised by both the slight differences between the present materials and those used in SPRinG, and the encouragement given to teachers to adapt materials to their pupils' needs. Primary school teachers have identified a need for detailed resources to compensate for their lack of confidence with the science curriculum (Harlen & Holroyd, 1997; Parker, 2004). The materials used in the present research, coupled with those utilised in the SPRinG KS2 project, can be regarded not only as supplying such resources for evaporation and condensation or force and motion, but also as providing a blue-print for the development of equivalent resources for other topics.

The second question addressed by the research was whether knowledge gains after the science programmes were dependent upon group work that displays the features highlighted by previous, less classroom-based investigations (e.g. Howe & Tolmie, 2003; Howe et al., 1995, 2000; Tolmie et al., 1993). The features of particular interest were those associated with pupil interaction, for full exploration of such features was beyond the scope of the SPRinG KS2 project. Foremost amongst the features flagged by earlier studies were occasions where pupils propose ideas and explain their reasoning, although disagreement, reference back to earlier material, and resolution of differences have also been pinpointed. The observational data collected here endorse the relevance of proposing and explaining when they occur during group work. Regression analyses demonstrated how knowledge gains after both science programmes were predicted by the proposition/explanation variable, but only in the context of group work: the variable had no relevance when computed for whole-class contexts. It would be in accordance with the background theory to suggest that in whole-class situations, pupils expect propositions and explanations to be evaluated by teachers, and this undermines their value for learning.

The data on disagreement, reference back and resolution are less straightforward than those relating to propositions and explanations. The frequency of disagreement was higher during group work than whole-class teaching, but disagreement was not itself consistently associated with post-test performance. On the other hand, disagreement was weakly loaded on the proposition/ explanation variable. It would be consistent with other research (e.g. Howe et al., 1995; Howe & McWilliam, 2006) to interpret this as indicating that disagreement creates a context where propositions and explanations are more likely, and by virtue of this provides indirect support to learning. As regards reference back and resolution, their lack of association with post-test performance may simply reflect their low frequency. Thus, the message of previous research, that these variables are potentially helpful but not essential, may still be valid. However, no matter what the potential, any benefits are likely to be minor compared with those stemming from propositions and explanations. Earlier, mention was made of the centrality of propositions and explanations to Piagetian thinking. It is interesting that the research has highlighted the importance of these variables, rather than variables that, from the Piagetian perspective, are relatively tangential.

Given the close association between observed interaction and post-test performance, it is perhaps surprising that the S-TOP ratings of group interaction were not similarly associated. The relevant scales did after all include ‘sharing and developing ideas’ and ‘use of exploratory talk’, with the concept of ‘exploratory talk’ focusing on explanation (Mercer, 2000). It is possible that the S-TOP approach was insufficiently discriminating, with the researchers obliged to treat each class as a single entity and therefore gloss over differences between pupils. The standard deviations in Table 3 indicate that the pupils varied in the extent to which they produced the dialogue features, and some of the variation may have been within, rather than between, classes. The whole-class nature of S-TOP also meant that sample size was reduced to 24, and this too may have compromised discrimination. From this perspective, it may be relevant that the beta values associated with the S-TOP group interaction scales were positive, and second only to the values associated with the adult role scales.

The association between the S-TOP adult role scales and post-test scores was of course both statistically significant and highly consistent with the emphasis in previous research on teacher monitoring and guidance. Blatchford et al. (2006) describe the ideal role as one where the teacher is ‘a guide on the side, not a sage on the stage’, and it is important that the present research endorses this. On the other hand, the research does not, on the face of it, endorse aspects of group work tapped by the S-TOP learning context or activities and tasks scales, for scores on these scales were not significantly associated with post-test performance. However, mean scores on both scales were almost at ceiling level (see Table 5), suggesting that most teachers made optimal arrangements. The skills training that preceded the science programmes emphasised group size and seating arrangements, and this may explain the lack of variation in the elements tapped by the learning context scales. As regards activities and tasks, they were of course orchestrated by the science materials themselves: as noted, the materials repeatedly instructed pupils to share ideas about how events should be explained, discuss their ideas and reach consensus, and record explanations in writing.

Although uncertainties remain, particularly surrounding some components tapped by S-TOP, it is clear that the second research question can also be answered in the affirmative. The success of the science programmes *was* reliant on group work that displays features highlighted by previous investigations. In particular, it required group work where teachers play a relatively 'hands-off' role, and pupils propose and justify ideas to other group members. From the favourable 'role of adults' ratings summarised in Table 5 and the relatively high levels of group-directed proposition and explanation presented in Table 3, the project must have been reasonably successful at fulfilling both requirements. Thus, one issue is what exactly it was about the project that allowed this to happen. To what extent, for instance, was it the skills training that preceded the science teaching, and to what extent was it the design of the science group tasks themselves? The skills training must have played some role: it was, after all, where the teachers were encouraged to guide rather than control. Nevertheless, it is clear that the science group tasks must also have contributed. As noted earlier, a small amount of group work occurred during the whole-class teaching. It took place around teacher-generated activities that were used to develop whole-class themes, and it was considerably less frequent than during the group sessions themselves. In fact, as the 'own group' figures in Table 3 show, group work in Classes 1 and 2 occurred with about 33% of the frequency of group work in Groups 1 and 2. However, propositions and explanations occurred during Classes 1 and 2 with only about 10% of the frequency of Groups 1 and 2, and when the skills training would have been relevant to all group contexts, this asymmetry must stem from the structure of the group tasks.

The role played by the group tasks needs to be remembered when returning, in conclusion, to the key issue of organisational principles. The critical message is that group work in elementary science should be organised to maximise the proposal and explanation of ideas in contexts where teachers are relatively non-directive. However, the potential for support from tasks that, like the present ones, are structured around sharing, discussion, agreeing and recording should not be over-looked. In highlighting the organisational principles, the present research endorses the relevance to elementary science of interaction in contexts of mutuality and equality, as emphasised in the background theories and research. Equally though, the research goes beyond existing material in at least two ways. First, in showing that productive interaction can occur in cross-age, as well as single-age, classrooms, it signals the power of interaction to *create* mutuality and equality in contexts where they cannot be presumed. This suggests a primacy for interaction, which is theoretically significant. Second, in examining principles that stem, in part at least, from experimental investigations (primarily Howe & Tolmie, 2003; Howe et al., 1995, 2000; Tolmie et al., 1993), the research helps to clarify the relation between experimental analysis and classroom practice. On the one hand, the research provides a relatively optimistic message, suggesting that experimental findings are not necessarily ‘subverted’ by local conditions and contextualised meanings (Engestrom, 2004; Kumpulainen et al., 2005). On the other hand, the research also signals a clear need for classroom-based work, and not only for testing viability in routine contexts. The results have, after all, streamlined the message from the experimental studies, by indicating, e.g., that references back and resolution may be even less central than previously thought. What is interesting, and certainly worthy of further analysis, is the suggestion via classroom-based research, that it is the features highlighted by *theory*, namely propositions and explanations, which have proved to be critical.

Acknowledgements

The reported research is part of a Scottish extension of the SPRinG project (http://www.tlrp.org/proj/phase11/phase2a.html), funded by the Economic and Social Research Council (ESRC) of Great Britain under its Teaching and Learning Research Programme. Thanks are due to the ESRC for their support, and to the SPRinG team (Ed Baines, Peter Blatchford, Maurice Galton, and Peter Kutnick) for their generous sharing of materials, data and relevant experience. Thanks are also due to the teachers, head-teachers and pupils from the participating schools for committing themselves to the project for most of the school year.

References

Association for Science Education. (2004). *Primary Science Review*. No. 83.

Bachmann, K., & Grossen, M. (2004). Explanations and modes of collaboration in tutor-tutee interactions at school. In K. Littleton, D. Miell, & D. Faulkner (Eds.), *Learning to collaborate, collaborating to learn* (pp. 111-132). New York: Nova Science.

Baines, E., Blatchford, P., & Chowne, A. (Under Review). Improving the effectiveness of pupil group work: Effects on attainment in science at KS2.

Baines, E., Blatchford, P., & Kutnick, P. (2003). Changes in grouping practices over primary and secondary school. *International Journal of Educational Research, 39,* 9-34.

Bennett, N., Desforges, C., Cockburn, A., & Wilkinson, B. (1984). *The quality of pupil learning experiences.* London: LEA.

Blatchford, P., Baines, E., Rubie-Davies, C., Bassett, P., & Chowne, A. (In Preparation). Improving the effectiveness of pupil group work: Effects on pupil-pupil and teacher-pupil interactions.

Blatchford, P., Kutnick, P., Baines, E., & Galton, M. (2003). Toward a social pedagogy of classroom group work. *International Journal of Educational Research, 39,* 153-172.

Christie, D., Thurston, A., Tolmie, A., Livingston, K., Topping, K., Howe, C., Jessiman, E., & Donaldson, C. (In Preparation). Supporting improvement in collaborative group work in urban and rural primary school classrooms.

Cohen, E.G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64,* 1-35.

Cohen, E.G., Lotan, R., & Leechor, C. (1989). Can classrooms learn? *Sociology of Education, 62*, 75-94.

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research, 5*, 331-343.

Department of Education and Science. (1989). *Science in the national curriculum*. London: HMSO.

Dewey, J. *Democracy and Education* (1916)

Ellis, S., & Rogoff, B. (1982). The strategies and efficacy of child versus adult teachers. *Child Development, 53*, 730-735.

Engestrom, Y. (2004). Towards a methodology of formative experiments: An activity-theoretical perspective on the study of teaching and learning. Paper presented at the Annual Conference of the Teaching and Learning Research Programme, Cardiff.

Forman, E. A., & Carr, N. (1992). Using peer collaboration to foster scientific thinking: What determines success? Paper presented at Annual Meeting of American Educational Research Association, San Francisco.

Galton, M.J., Simon, B., & Croll, P. (1980). *Inside the primary classroom.* London: Routledge & Kegan Paul.

Galton, M.J., Hargreaves, L., Comber, C., Wall, D., & Pell, A. (1999). *Inside the primary classroom: 20 years on.* London: Routledge.

General Register Office for Scotland. (2004). *2001 Census results.* Retrieved from www.gro-scotland.gov.uk/grosweb/grosweb.nsf/pages/censushm.

Harlen, W., & Holroyd, C. (1997). Primary teachers' understanding of concepts in science: Impact on confidence and teaching. *International Journal of Science Education, 19,* 93-105.

Harlen, W., & Qualter, A. (2004). *The teaching of science in primary schools. 4th edition.* London: David Fulton.

Hennessy, S., Twigger, D., Driver, R., O'Shea, T., O'Malley, C.E., Byard, M., Draper, S., Hartley, R., Mohamed, R., & Scanlon, E. (1995). A classroom intervention using a computer-augmented curriculum for mechanics. *International Journal of Science Education, 17,* 189-206.

Howe, C., & McWilliam, D. (In Press). Opposition in social interaction between children: Why intellectual benefits do not mean social costs. To appear in *Social Development.*

Howe, C., McWilliam, D., & Cross, G. (2005). Chance favours only the prepared mind: Incubation and the delayed effects of peer collaboration. *British Journal of Psychology, 96,* 67-93.

Howe, C.J., Rodgers, C., & Tolmie, A. (1990). Physics in the primary school: Peer interaction and the understanding of floating and sinking. *European Journal of Psychology of Education, V,* 459-475.

Howe, C.J., & Tolmie, A. (2003). Group work in primary school science: Discussion, consensus and guidance from experts. *International Journal of Educational Research, 39,* 51-72.

Howe, C.J., Tolmie, A., Duchak-Tanner, V., & Rattray, C. (2000). Hypothesis testing in science: Group consensus and the acquisition of conceptual and procedural knowledge. *Learning and Instruction, 10,* 361-391.

Howe, C.J., Tolmie, A., Greer, K., & Mackenzie, M. (1995). Peer collaboration and conceptual growth in physics: Task influences on children's understanding of heating and cooling. *Cognition and Instruction, 13*, 483-503.

Howe, C.J., Tolmie, A., & Rodgers, C. (1992). The acquisition of conceptual knowledge in science by primary school children: Group interaction and the understanding of motion down an inclined plane. *British Journal of Developmental Psychology, 10*, 113-130.

Jaques, D. (2000). *Learning in groups. 3rd edition*. London: Kogan Page.

Kruger, A., & Tomasello, M. (1986). Transactive discussions with peers and adults. *Developmental Psychology, 22*, 681-685.

Kumpulainen, K., Kangassalo, M., & Vasama, S. (2005). Studying the (con)textuality of explanations in the learning of science. Paper presented at the 11th Biennial Conference of EARLI, Nicosia, Cyprus.

Learning and Teaching Scotland. (2000). *Environmental studies: Science.* Edinburgh: Scottish Executive.

Lou, Y., Abrami, P.C., Spence, J.C., Poulsen, C., Chambers, B., & d’Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*, 423-458.

Mercer, N. (2000). *Words and minds: How we use language to think together*. London: Routledge.

Miell, D., & MacDonald, R. (2000). Children's creative collaborations: The importance of friendship when working together on a musical composition. *Social Development, 9*, 348-369.

Miller, S., & Brownell, C. (1975). Peers, persuasion and Piaget: Dyadic interaction between conservers and non-conservers. *Child Development, 46*, 992-997.

Morgan, L., Hargreaves, D., & Joiner, R. (2000). Children's collaborative music composition: Communication through music. In R. Joiner, K. Littleton, D. Faulkner, & D. Miell (Eds.), *Rethinking collaborative learning* (pp. 52-64). London: Free Association.

Mugny, G., & Doise, W. (1978). Socio-cognitive conflict and structure of individual and collective performances. *European Journal of Social Psychology, 8*, 181-192.

Parker, J. (2004). The synthesis of subject and pedagogy for effective learning and teaching in primary science education*. British Educational Research Journal, 30*, 819-839.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context.* Oxford: Oxford University Press.

Slavin, R.E. (1995). *Co-operative learning: Theory, research and practice. 2nd edition*. Boston: Allyn & Bacon.

Tolmie, A., Howe, C.J., Mackenzie, M., & Greer K. (1993). Task design as an influence on dialogue and learning: Primary school group work with object flotation. *Social Development, 2*, 183-201.

Topping, K., Tolmie, A., Christie, D., Donaldson, C., Howe, C., Jessiman, E., Livingston, K., & Thurston, A. (In Preparation). Social effects of collaborative learning in elementary schools.

Ward, H., Roden, J. Hewlett, C., & Foreman, J. (2005). *Teaching science in the primary school*. London: Paul Chapman.

Webb, N. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research, 13*, 21-39.

Williams, J.M., & Binnie, L.M. (2002). Children's concepts of illness: An intervention to improve knowledge. *British Journal of Health Psychology*, *7*, 129-147.

Williams, J., & Tolmie, A. (2000). Conceptual change in biology: Group interaction and the understanding of inheritance. *British Journal of Developmental Psychology*, *18*, 625-649.

Table 1: Sample Characteristics

Intervention/ Control Classes	School Location	Composite Classes	Single-Age Classes
Intervention	Urban	6 schools (N = 135/143)	6 schools (N = 141/114)
Intervention	Rural	6 schools (N = 117/92)	6 schools (N = 121/116)
Control	Urban	-	2 schools (N = 48/40)
Control	Rural	-	1 school (N = 23/20)

Note: The bracketed figures in the two right-hand columns show the numbers of pupils supplying pre- and post-test data, and therefore included in the formal analyses. The first figure relates to the evaporation and condensation data, and the second figure relates to the force and motion, e.g. (N = 135/143) means that 135 pupils supplied pre- and post-test data for evaporation and condensation, and 143 did this for force and motion.

Table 2: Mean Pre-test and Post-test Scores (SD in Brackets)

	Intervention					Control
	Single-Age		Composite		Total Intervention	
	Urban	Rural	Urban	Rural		
Evaporation and Condensation						
Pre-test	8.35 (2.70)	9.45 (3.02)	9.85 (5.07)	9.26 (2.73)	9.21 (3.59)	10.30 (3.26)
Post-test	11.32 (3.60)	12.37 (3.41)	12.67 (3.17)	12.48 (3.98)	12.18 (3.57)	10.14 (3.00)
Force and Motion						
Pre-test	19.85 (4.97)	22.47 (4.69)	19.87 (4.47)	19.86 (5.12)	20.51 (4.90)	23.15 (5.09)
Post-test	23.04 (5.28)	24.86 (5.04)	22.78 (5.19)	23.59 (4.94)	23.52 (5.18)	23.88 (5.04)

Table 3: Mean Frequency of Classroom Observation Variables in Whole-Class and Small-Group Settings (SD in Brackets)

	Whole-Class		Small-Group		F Value
	Class 1	Class 2	Group 1	Group 2	(*df*=3,369)
Social Context					
Alone	5.24 (2.61)	5.67 (2.62)	0.55 (1.47)	0.46 (1.42)	233.48 ***
Teacher	0.41 (0.76)	0.37 (0.92)	0.56 (1.11)	0.40 (0.68)	*ns*
Own Group	2.09 (2.38)	1.90 (2.34)	6.81 (1.79)	7.01 (1.79)	231.12 ***
Other Group	0.23 (0.72)	0.10 (0.35)	0.32 (1.07)	0.16 (0.43)	*ns*
Dialogue					
Proposition	0.27 (0.57)	0.34 (0.63)	2.26 (1.93)	3.09 (2.51)	112.58 ***
Disagreement	0.08 (0.30)	0.03 (0.28)	0.48 (0.82)	0.50 (0.74)	23.11 ***
Explanation	0.15 (0.38)	0.12 (0.39)	0.91 (1.14)	1.18 (1.50)	40.61 ***
Reference Back	0 (N/A)	0.03 (0.18)	0.08 (0.35)	0.10 (0.30)	4.23 *
Resolution	0 (N/A)	0 (N/A)	0.01 (0.09)	0.03 (0.18)	*ns*
Instruction	0.09 (0.34)	0.12 (0.45)	0.57 (0.91)	1.13 (1.23)	46.03 ***
Question	0 (N/A)	0 (N/A)	0.09 (0.34)	0.08 (0.27)	6.72 **
Uncodable	2.08 (1.86)	1.64 (1.58)	2.72 (1.67)	2.16 (1.70)	*ns*

* p<.05, ** p<.01, *** p<.001

Table 4: Correlations between Group Dialogue Frequencies and Post-test Scores

Dialogue Variable	Evaporation and Condensation Post-test	Force and Motion Post-test
Proposition/Explanation	.27 **	.26 **
Disagreement (Group 1)	.18 *	.17
Disagreement (Group 2)	.12	.14
References Back (Group 1)	.09	.02
References Back (Group 2)	–.05	.03
Resolution (Group 1)	.12	.12
Resolution (Group 2)	–.08	.01
Instruction (Group 1)	.08	.04
Instruction (Group 2)	.14	.21 *
Question (Group 1)	.04	.13
Question (Group 2)	.16	.07

* p<.05, ** p<.01

Table 5: Analyses of S-TOP Ratings

	Group 1 Mean (SD)	Group 2 Mean (SD)	Correlation Group 1 vs Group 2	Regression Evaporation and Condensation	Regression Force and Motion
Learning Context (Max = 12)	11.81 (0.53)	11.68 (0.78)	.49 ***	β = .16, *ns*	β = .32, *ns*
Activities and Tasks (Max = 21)	19.66 (2.34)	19.99 (2.61)	.18 *	β = .21, *ns*	β = .29, *ns*
Role of Adults (Max = 27)	23.17 (3.07)	23.40 (2.94)	.47 ***	β = .50, t = 2.70*	β = .64, t = 3.80***
Group Interaction (Max = 33)	28.06 (3.29)	31.02 (2.85)	.40 ***	β = .32, *ns*	β = .34, *ns*

* p<.05, *** p<.001