This article was downloaded by: [University of Colorado - Health Science Library]
On: 13 December 2014, At: 13:17
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

International Journal of Science Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tsed20

Principled Improvement in Science: Forces and proportional relations in early secondary-school teaching

Christine Howe[a], Sonia Ilie[a], Paula Guardia[b], Riikka Hofmann[a], Neil Mercer[a] & Fran Riga[a]

[a] Faculty of Education, University of Cambridge, Cambridge, UK

[b] Facultad de Educación, Pontificia Universidad Católica de Chile, Santiago, Chile

Published online: 07 Nov 2014.

To cite this article: Christine Howe, Sonia Ilie, Paula Guardia, Riikka Hofmann, Neil Mercer & Fran Riga (2015) Principled Improvement in Science: Forces and proportional relations in early secondary-school teaching, International Journal of Science Education, 37:1, 162-184, DOI: 10.1080/09500693.2014.975168

To link to this article: http://dx.doi.org/10.1080/09500693.2014.975168

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms &

Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

International Journal of Science Education, 2015
Vol. 37, No. 1, 162–184, http://dx.doi.org/10.1080/09500693.2014.975168

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Principled Improvement in Science: Forces and proportional relations in early secondary-school teaching

Christine Howe[a*], Sonia Ilie[a], Paula Guardia[b], Riikka Hofmann[a], Neil Mercer[a] and Fran Riga[a]
[a]*Faculty of Education, University of Cambridge, Cambridge, UK;* [b]*Facultad de Educación, Pontificia Universidad Católica de Chile, Santiago, Chile*

In response to continuing concerns about student attainment and participation in science and mathematics, the *epiSTEMe* project took a novel approach to pedagogy in these two disciplines. Using principles identified as effective in the research literature (and combining these in a fashion not previously attempted), the project developed topic modules for early secondary-school teaching in the UK, arranged for their implementation in classrooms, and evaluated the results. This paper reports the development, implementation, and evaluation of one of the *epiSTEMe* science modules. Entitled *Forces and Proportional Relations*, the module covers standard curricular material in the domain of forces, while paying particular attention to the proportional nature of many key constructs. It was developed in collaboration with a small group of teachers; implemented subsequently in 16 classrooms, in all cases involving students from the first year of secondary school; and evaluated through comparison with first-year students in 13 control classrooms who were studying the topic using established methods. Evaluation addressed topic mastery and opinions about the topic and the manner in which it was taught. While further research is required before definite conclusions are warranted, results relating to topic mastery provide grounds for optimism about the *epiSTEMe* approach. Furthermore, student opinions about the module were positive.

Keywords: *Science education; Early secondary-school; Dialogic teaching; Forces; Proportional reasoning*

Introduction

The effectiveness of school education in science and mathematics is an enduring concern in many countries (Gilbert, 2006; Kilpatrick, Swafford, & Findell, 2001;

*Corresponding author. Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, UK. Email: cjh82@cam.ac.uk

© 2014 Taylor & Francis

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Roberts, 2002), with levels of attainment and post-compulsory participation both spotlighted as significant issues. In the UK, concerns about attainment are often expressed with reference to the nation's middling and relatively static position in international league tables (e.g. Programme for International Student Assessment, 2009, 2012; Trends in International Mathematics and Science Study, 1995, 2011). With post-compulsory participation, the focus is sometimes upon proportionate continuation amongst the relevant population as a whole and sometimes upon continuation amongst identifiable subgroups. For instance, a recent 'state of the nation report' on post-compulsory schooling (Royal Society, 2011) highlights with concern that only 17.3% of 'the potentially eligible population' in the UK is enrolled for core sciences and/or mathematics. At the same time, about 65% of the students taking A-level mathematics are boys as are around 80% of the students taking A-level physics (Department for Education and Skills, 2007), even though girls have for some time performed at least as well as boys in these subjects at earlier stages. (A-level is the main post-compulsory academic qualification in most parts of the UK.)

With this situation as backcloth, two key UK organizations, the *Economic and Social Research Council* and the *Institute of Physics*, launched a programme of research dedicated to identifying factors relevant to attainment and participation in science and mathematics. The programme, which came to be known as the *Targeted Initiative on Science and Mathematics Education*, comprised five projects of which one, styled *epiSTEMe* for *Effecting principled improvement in STEM education*, is the focus here. The starting point for *epiSTEMe* was the copious evidence that classroom processes play a very large role in determining student outcomes (e.g. Hattie, 2012). The project was, in part, an attempt to use research literature to identify optimal processes, and to distil conclusions in pedagogical principles for science and mathematics. It was in addition an attempt to implement the principles during the early years of what, in the UK, is termed 'secondary education', and to evaluate the impact. With the focus upon the early years of secondary school, evaluation inevitably revolved around attainment rather than post-compulsory participation. However, it included attitudinal measures as well as indices of attainment, and the implications for student subgroups were examined, including for those whose participation is a concern. The present paper outlines implementation of principles and evaluation of outcomes in one of the several contexts that *epiSTEMe* covered, and reports some results.

General Pedagogic Principles

The general principles underpinning *epiSTEMe* are detailed in Ruthven et al. (2011). In brief, one source of these principles was the North American programmes that have been judged 'exemplary' on the basis of evidence for their effectiveness in multiple sites for multiple samples (Department of Education, 1999). It is clear that many of these programmes are organized around carefully structured problem situations, which are designed to appeal to students' wider experiences and to inculcate ideas of acting as scientists/mathematicians in developing key disciplinary ideas (Bransford, Brown, & Cocking, 2000; Duschl, Schweingruber, & Shouse, 2007; Kilpatrick et al., 2001). Accordingly, *epiSTEMe* deployed series of problems in science and mathematics, which embodied these features. Problems were represented on PowerPoint slides for classroom usage, and supported with notes for teachers and technicians and workbooks for students.

Equally, *epiSTEMe* was strongly influenced by UK research on classroom interaction (Alexander, 2008; Howe et al., 2007; Mercer, Dawes, Wegerif, & Sams, 2004; Mercer & Sams, 2006; Scott, Mortimer, & Aguiar, 2006), which highlights the value of small-group and whole-class discussion in which students talk in an exploratory fashion, explaining their ideas and using reasoned argument to consider different perspectives. This form of interaction, which has come to be known as 'dialogic teaching', has been used successfully to promote 'thinking together' in science (Mercer et al., 2004) and mathematics (Mercer & Sams, 2006), with students using talk effectively as the tool for reasoning, and showing enhanced reasoning, understanding, and problem-solving within the two disciplines. At the same time, dialogic teaching is not straightforward to implement, posing challenges for both students and teachers. As a result, *epiSTEMe* provided support in establishing a dialogic style prior to deployment with the targeted disciplinary content while also presenting content in a fashion that scaffolded the intended style. In particular, teachers were invited to professional development sessions where target practices were discussed and illustrated through video-extracts. They were given activities to try with their students, which included formulating 'ground rules' for effective talk, practising skills with tasks that paralleled those that would feature in the disciplinary teaching, and scoring talk for, for example, 'listening carefully' and 'explaining ideas'.

Principled Approach to Forces

The general principles were applied to teaching 'modules' relating to four areas that are addressed in the early secondary curriculum, two areas from science and two from mathematics. This paper focuses on the module relating to forces, one of the science areas. The topic was included in *epiSTEMe* because it is a central component of all science curricula in the UK (e.g. Education Scotland, 2011; Qualifications and Curriculum Authority, 2007), yet it is known to be extremely challenging for students. It is perhaps the single most fertile source of the misconceptions in science that have dominated the research literature for over 30 years, and that have been shown to endure into adulthood despite extensive teaching (see e.g. White & Gunstone's (2008) analysis of the 8000+ entries in Duit's (2007) bibliography). There are undoubtedly many reasons for the difficulties. For instance, key ideas relating to forces can be counter-intuitive: in a world where friction is ubiquitous but scarcely noticed, it is far from obvious that moving objects continue to move with constant velocity unless subject to force. In a world where media images depict backwards fall from inside moving carriers, it is also far from obvious that when viewed from outside the carriers objects actually fall forwards. The need to combat counter-intuitiveness is both recognized implicitly in the *epiSTEMe* principle of appealing to

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

student experiences and a challenge when envisaging how the principle should be implemented in practice.

Less widely discussed is the fact that many pivotal and challenging concepts relating to forces depend upon a grasp of proportionality. For example, density is directly proportional to mass and inversely proportional to volume, speed is directly proportional to distance travelled and inversely proportional to time taken, and stretch is directly proportional to suspended load and inversely proportional to number of springs. As noted in Howe, Nunes, Bryant, Bell, and Desli (2010), the proportionality inherent in such concepts is often sidestepped in science education: density, speed, and stretch are typically analysed without reference to their constituent quantities. However, as a consequence, something crucial may be missed, for the difficulties that students of all ages experience with proportional reasoning have been widely documented (e.g. Boom, Hoijtink, & Kunnen, 2001; Dean & Frankhouser, 1988; Howe, Nunes, & Bryant, 2010; Lesh, Post, & Behr, 1988; Piaget, Grize, Szeminska, & Bang, 1977; Piaget & Inhelder, 1975; Siegler, 1981). Furthermore, it has become apparent that mathematical representation of proportional relations can help to overcome the difficulties (Schwartz, Martin, & Pfaffman, 2005; Schwartz & Moore, 1998), yet science in the UK is often pointedly non-mathematical in the early secondary years. Qualitative grasp is emphasized rather than quantification. This is despite pressure to bring science and mathematics into closer alignment throughout the school years (e.g. Department for Education and Employment, 2001; SCORE, 2011).

In planning the module on forces, it was felt that a more explicit (and explicitly mathematical) treatment of proportionality might pay dividends. Thus, this became a further design principle underpinning the module, along with the broader principles outlined earlier that guided *epiSTEMe* in general. The module was, as a result, called *Forces and proportional relations* (although for brevity it will be referred to here as *Forces*). The section that follows describes the module in detail, showing how the design principles operated jointly to produce a coherent programme for classroom teaching. Later sections outline how the module was implemented in UK schools, and report the main findings from an evaluation study in which the impact of the *Forces* module on student learning and attitudes was compared with the impact of established methods of teaching. The teaching materials associated with the module have been translated into Spanish, and the UK implementation and evaluation have been replicated with minor modification in 18 schools in Chile. However, because the Chilean work is reported elsewhere (e.g. Larrain, 2013; Larrain, Howe, & Freire, in press), the following is largely restricted to what happened within the UK. Chilean results are referred to only where they help to clarify the UK data.

Module Development

Module Structure

The *Forces* module was developed over a two-year period in collaboration with a core group of four teachers from local secondary schools, who attended eight full-day

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

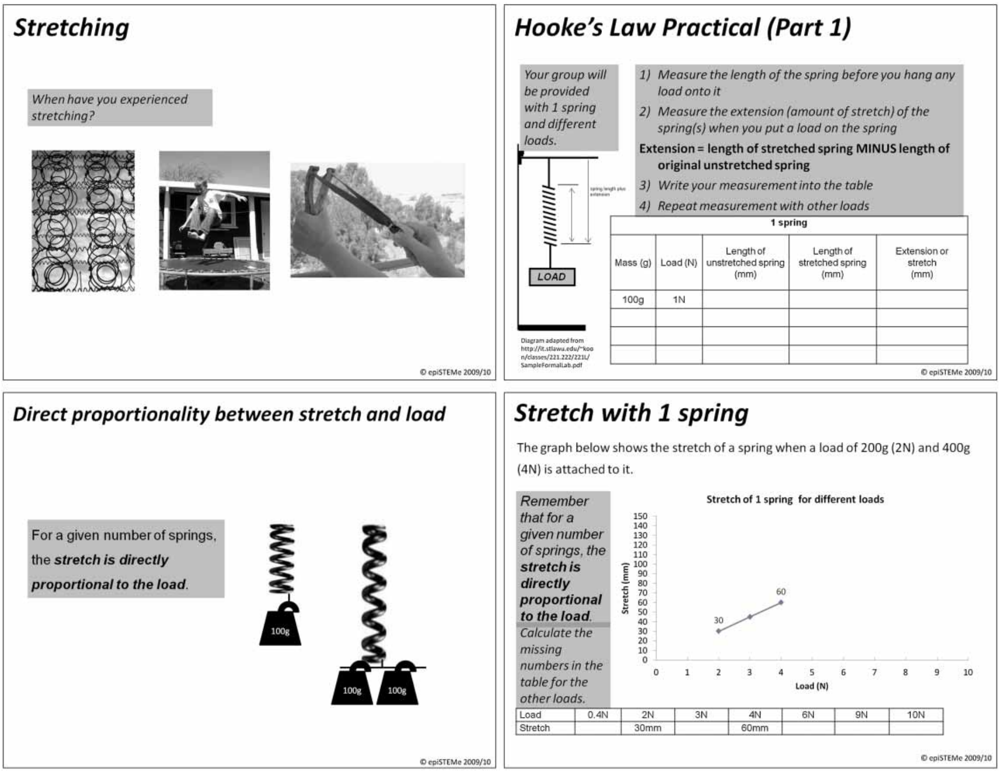
Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

workshops at the university where the research team was based. Further teachers participated in a subset of the workshops. Topic selection was guided via curriculum requirements and the teachers' views about what is critical for teaching forces during the first year of secondary education. Materials that the teachers supplied were embellished in accordance with *epiSTEMe* principles, and incorporated within the module. In its final form, the module comprised 10 core 'lessons' plus optional extensions and homework: (1) Introduction; (2) Balanced and unbalanced forces; (3) Measuring forces; (4) Stretching (two lessons); (5) Flotation and density (two lessons); (6) Surface friction; (7) Stopping distance (two lessons). Lessons were notionally each of 50-minute duration, but were divided into three or four parts to facilitate adjustment to teaching periods of differing lengths. The module was detailed in 71-page teaching notes supplemented with notes for laboratory technicians, together with PowerPoint slides and student booklets for classroom presentation. All materials are available at https://camtools.cam.ac.uk/wiki/site/6f837af4-d690-45ac-001a-7d377cf9cf3f/forces%20module.html.

All topics reflected the general *epiSTEMe* principle of achieving relevance to student experiences and interests, and wherever appropriate proportionality was highlighted and mathematical representation implemented. For instance, the activities comprising the two lessons on stretching began with the identification of everyday contexts where stretching occurs, and the viewing and discussion of a video on bungee jumping. This was followed with a straightforward practical exercise involving the measurement of spring stretch (extension) when different loads are suspended from a single spring. With the results tabulated and graphed, the relation between stretch and load was depicted in terms of direct proportionality, and stretch with further loads was to be predicted using mathematics (see Figure 1 for four of the six PowerPoint slides provided to support these activities). The inversely proportional relation between stretch and number of springs was then introduced in a parallel fashion, that is, a practical exercise where load was constant and springs varied was followed with tabulation, graphing, specification of the relation, and mathematically supported prediction. The lessons concluded with bungee jumping revisited in the light of the preceding activities.

All topics also reflected the general *epiSTEMe* principle of supporting high-quality dialogue in small-group and whole-class settings. For instance, the lesson on balanced and unbalanced forces contained several activities where everyday scenarios were depicted and statements presented that covered the correct analysis together with what the literature indicates are common misconceptions (see the top half of Figure 2 for two examples). After deciding independently whether they agreed or disagreed with each statement, students were to discuss the statements in small groups and collectively identify the correct one. Group decisions and the reasoning behind decisions were to be collated in whole-class plenary sessions, and differences between groups (over decisions and reasoning) were to be discussed and resolved. While slides were provided for teachers to use in supporting student understanding (see the bottom half of Figure 2), the emphasis was upon guiding towards appropriate resolution of differences rather than imposing correct

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014



Stretching

When have you experienced stretching?

© epiSTEMe 2009/10

Hooke's Law Practical (Part 1)

Your group will be provided with 1 spring and different loads.

1) *Measure the length of the spring before you hang any load onto it*
2) *Measure the extension (amount of stretch) of the spring(s) when you put a load on the spring*

Extension = length of stretched spring MINUS length of original unstretched spring

3) *Write your measurement into the table*
4) *Repeat measurement with other loads*

1 spring				
Mass (g)	Load (N)	Length of unstretched spring (mm)	Length of stretched spring (mm)	Extension or stretch (mm)
100g	1N			

Diagram adapted from http://it.stlawu.edu/~koon/classes/221.222/221L/SampleFormalLab.pdf

© epiSTEMe 2009/10

Direct proportionality between stretch and load

For a given number of springs, the ***stretch is directly proportional to the load***.

© epiSTEMe 2009/10

Stretch with 1 spring

The graph below shows the stretch of a spring when a load of 200g (2N) and 400g (4N) is attached to it.

Remember that for a given number of springs, the ***stretch is directly proportional to the load****.* *Calculate the missing numbers in the table for the other loads.*

Load	0.4N	2N	3N	4N	6N	9N	10N
Stretch		30mm		60mm			

© epiSTEMe 2009/10

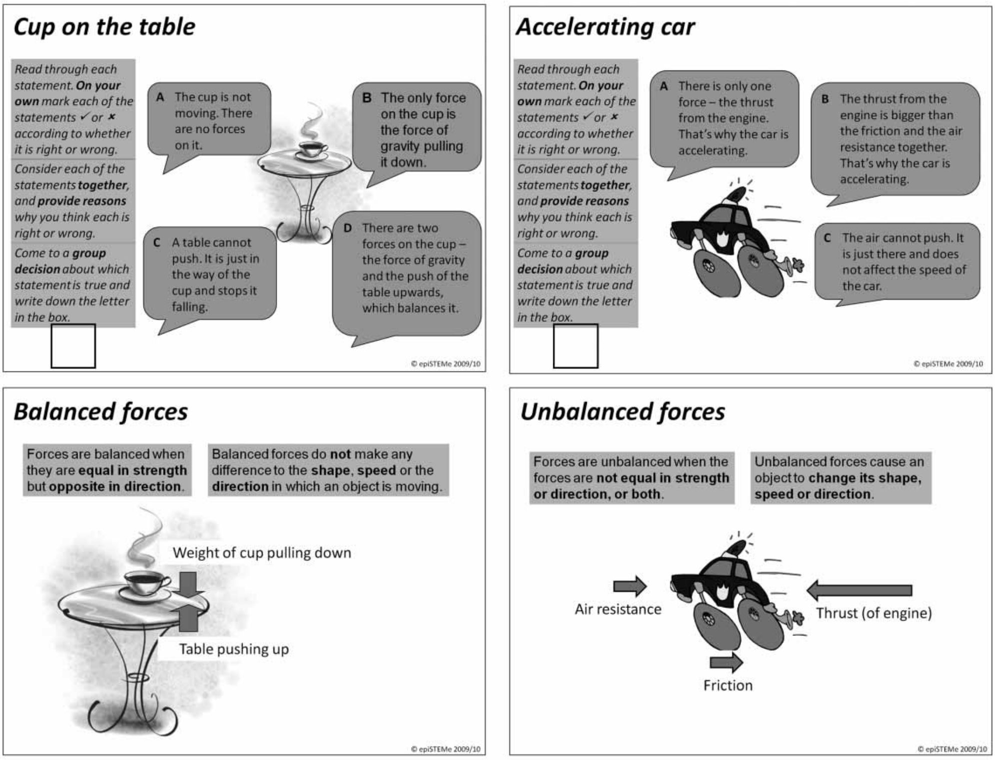
Figure 1. Sample of slides used in lessons on stretching

solutions. Thus, the need for what Scott et al. (2006) term the 'authoritative' voice was recognized, but its introduction was expected to lie within the dialogic process.

Evaluation Instruments

Concurrently with module design, three instruments were developed to evaluate its effectiveness: (1) Observation schedule; (2) Opinion questionnaire; (3) Knowledge tests. The first two instruments were generic across the *epiSTEMe* modules, and will be detailed in Ruthven et al. (in press). In brief, the observation schedule comprised nine categories that are central to the concept of dialogic teaching, for example, 'Teacher asks for explanation, clarification or reasoning', 'Teacher draws out differences between student ideas', 'Student gives reason', 'Different perspectives are discussed for at least one minute'. The intention was that a researcher would observe for successive 6-minute periods across full lessons, and indicate on checklists the categories that occurred within each period. With reference to category frequency, an index would be obtained of how successfully the dialogic component had been implemented. Before using the schedule in schools, the researcher and a colleague

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014



Cup on the table

Read through each statement. ***On your own*** *mark each of the statements ✓ or ✗ according to whether it is right or wrong.*

Consider each of the statements ***together****, and* ***provide reasons*** *why you think each is right or wrong.*

Come to a ***group decision*** *about which statement is true and write down the letter in the box.*

□

A The cup is not moving. There are no forces on it.

B The only force on the cup is the force of gravity pulling it down.

C A table cannot push. It is just in the way of the cup and stops it falling.

D There are two forces on the cup – the force of gravity and the push of the table upwards, which balances it.

© epiSTEMe 2009/10

Accelerating car

Read through each statement. ***On your own*** *mark each of the statements ✓ or ✗ according to whether it is right or wrong.*

Consider each of the statements ***together****, and* ***provide reasons*** *why you think each is right or wrong.*

Come to a ***group decision*** *about which statement is true and write down the letter in the box.*

□

A There is only one force – the thrust from the engine. That's why the car is accelerating.

B The thrust from the engine is bigger than the friction and the air resistance together. That's why the car is accelerating.

C The air cannot push. It is just there and does not affect the speed of the car.

© epiSTEMe 2009/10

Balanced forces

Forces are balanced when they are **equal in strength** but **opposite in direction**.

Balanced forces do **not** make any difference to the **shape**, **speed** or the **direction** in which an object is moving.

© epiSTEMe 2009/10

Unbalanced forces

Forces are unbalanced when the forces are **not equal in strength or direction, or both**.

Unbalanced forces cause an object to **change its shape, speed or direction**.

© epiSTEMe 2009/10

Figure 2. Sample of slides from lesson on balanced and unbalanced forces

independently coded videotapes obtained during the project's first two years, with average inter-judge agreement across categories of 72%.

The opinion questionnaire was to be administered in the final lesson and comprised 10 pairs of 7-point Likert scales covering views about the module: (1) Pitch, for example, 'These lessons were too difficult for me'; (2) Interest, for example, 'Learning about this topic has been interesting'; (3) Effort, for example, 'I worked hard in the lessons on this topic'; (4) Thinking, for example, 'These lessons have made me think a lot'; (5) Future, for example, 'I hope we don't study this topic again'; (6) Learning, for example, 'These lessons have taught me a lot about this topic'; (7) Understanding, for example, 'These lessons have helped me make sense of this topic'; (8) Explanation, for example, 'I've got better at explaining things through taking part in these lessons'; (9) Value, for example, 'These lessons have shown me why it's important to study this topic'; (10) Application, for example, 'These lessons have helped me see how this topic applies to real life'. The questionnaire was printed on a single A4 sheet for written completion.

The knowledge tests were specific to the *Forces* module and began with items where correct answers had to be selected from options, for example, 'What is the weight of a 20 kg box on the Earth?' Options: 2N, 20N, 200N; 'Adding tap water to salt water' Options: increases the density of the solution, decreases the density of the solution,

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

does not affect the density of the solution (maximum score for the section = 6). These were followed with items relating to force diagrams, for example, the four forces acting on a flying aeroplane (shown via labelled arrows) were associated with questions like 'Which arrow represents air resistance?' 'When the plane is flying at a constant height, which two forces must be balanced?' (maximum section score = 5). The next set of items required the calculation of weight on other planets and in deep space, and a brief statement of why weight on the planet differs from weight on Earth (maximum section score = 3). Addressing flotation and density, some items in the fourth set requested brief statements, for example, 'Ian floats on water [in the swimming pool]. Why does Ian float on the water even though gravity is pulling him down?' while other items involved selection from options, for example, 'Ian swims wearing his clothes. How are the forces different from when he swims wearing his swimming costume? Tick two boxes' Options: Weight increases, gravity increases, friction increases, friction decreases (Maximum section score = 6). The concluding items presented three speeds (e.g. 20, 25, 30 miles/hour) and associated stopping distances, showed the thinking and braking distances for one speed, and requested calculation of one or both of these distances for the other speeds. Two influences on actual stopping distance then had to be nominated, with brief statements to be provided of how they operate (Maximum section score = 5). In total then, the maximum score for each test was 25.

There were three knowledge tests: (1) a pre-test, to be administered during the first lesson to assess initial understanding; (2) an immediate post-test, to be administered during the final lesson to assess understanding upon completion of teaching; (3) a deferred post-test, to be administered about one month after the immediate post-test, to assess long-term gain. All were presented in booklets in which responses were to be written. Every item that appeared in the pre-test appeared in one of the post-tests, with 50% overlap between the pre-test and immediate post-test and 59% overlap between the pre-test and deferred post-test. All items had been extracted (with minor modification) from previous national assessments http://www.satspapers.co.uk/sats-papers/ks3/science and given the *Forces* module's grounding in the statutory curriculum and the involvement of practising teachers in its design can be assumed to be at an appropriate level for early secondary-school students, whether taught via *epiSTEMe* or not. At the same time, items were chosen to probe areas where profound misconceptions have been identified. For instance, an average of 60% of the score for each test depended upon appropriate use and/or definition of gravity, surface friction, air and water resistance, upthrust and density, and speed and speed change. Students would not have performed well holding the commonplace beliefs (Howe, Taylor Tavares, & Devine, 2012, 2014) that stopping depends on expended impetus or physical obstacles, fluids suck objects down, or objects fall through air with constant or decelerating velocity.

Test items were selected from a much larger battery that was administered during the two years of module development to over 100 students from first-year secondary classes. These students' proportionate accuracy per item provided indices of item difficulty. Test items were selected such that the pre-test, immediate post-test, and deferred post-test were of equivalent difficulty when indices were averaged across constituent items. Towards the end of the development period, five of the secondary teachers who had been involved with design implemented the module with first-year students and administered the tests to the planned schedule (i.e. start-of-module, end-of-module, after one month). Encouragingly the average effect size for student gain after *epiSTEMe* was +0.82 (Cohen's *d*) for pre- to immediate change and +0.76 for pre- to deferred change. This compared with values of +0.23 and +0.18, respectively, from students in an informal comparison group, that is, first-year students from the same schools whose teachers were not involved with *epiSTEMe* and therefore covered forces using established methods. The *epiSTEMe* teachers were both surprised and impressed at how well their students had responded to the mathematical elements. As one teacher explained:

> We wrote on the board that thinking distance was 6 m and braking distance was 6 m, and then I asked them to predict what they would be with a speed of 40 mph—everyone thought 12 and 12. So then we tried it—and of course found that braking distance was greater. But they were able to come up with really good explanations as to why thinking distance doubled. Someone realized that braking distance quadrupled—I struggle to get my Year 11s [fifth-year] to notice that! And someone else gave a really good explanation as to why braking distance more than doubled. Overall, we hardly did any writing in the booklets but it was a fantastic lesson and I was really chuffed with how well they were working and talking together and they REALLY enjoyed it!

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Large-Scale Implementation

Module Implementation

With design work completed and some encouraging albeit preliminary results, the module was ready for large-scale implementation and formal evaluation. To this end, an approach was made to all secondary schools within a c.50-mile radius of the university (and therefore limited to England rather than the UK as a whole). Information was provided about the project, together with an invitation to nominate up to two science teachers who were involved with first-year classes (Year 7 in England). No specific guidance was given to schools over the selection of teachers, apart from recommending that only a minority of students in their Year 7 class(es) should fall amongst the lowest 20% in terms of end-of-primary-school attainment. Nominated teachers were invited to a briefing session held in the university, where a broad overview was provided of what participation in the project would involve. It was emphasized that only 50% of participants would be given the *Forces* module during the following year, that is, constitute the intervention group. The remainder would be asked to act as a control group, which would involve teaching forces via established methods, while administering the *epiSTEMe* evaluation instruments. It was stressed that after a 12-month interval, control teachers would receive module materials and be offered support in their implementation that was equivalent to the intervention group. Schools whose teachers agreed to proceed on this basis were listed in order of their most recent CVA2-4 score (in England, CVA2-4 is one of several nationwide and standardized indices of efficacy—see Department for Education, 2010). Using the list and in the hope of achieving cross-condition equivalence, adjacent pairs of schools were assigned randomly to the intervention and control conditions. Thus, when two teachers had volunteered from a single school, they were both placed in the same condition.

Implementation was supported through two full-day professional development sessions for teachers, intentionally mimicking the amount of support that is typically provided in England for educational innovation and therefore permitting assessment of the module's effectiveness under normal circumstances. The first session (which took place before, or early in, the school year) was primarily devoted to introducing the principles of dialogic teaching. As mentioned earlier, introduction relied heavily on video-extracts, and many of these extracts involved the teachers who contributed to module design, recorded as they tried activities out with their students. It also covered the preparatory classroom activities alluded to earlier that embodied the dialogic principles, that is, activities that involved formulating ground rules, scoring talk for effectiveness, and practising discussion tasks. One example of the latter was evaluating statements akin to those in Figure 2 on the theme of 'Are nurses scientists?' Teachers were encouraged to use the activities with the Year 7 class they would be involving in *epiSTEMe* prior to the next professional development session. Towards the close of the first session teachers in the intervention group were given copies of an attitude questionnaire (see Ruthven et al., in press, for details), which they were asked to administer to students in their Year 7 class. The purpose of the questionnaire was to establish pre-existing attitudes to science that could be taken into account when assessing opinions about the module and learning gains. The questionnaire comprised 20 7-point Likert scales, with five scales relating to each of the following: (1) Ability in science, for example, 'I'm good at science'; (2) Enthusiasm for science, for example, 'Science is boring'; (3) Prospective involvement in science, for example, 'I'd like a job that involves using science'; (4) Wider value of science, for example, 'Everybody will need to know some science in their adult life'. Teachers in the control group were sent the attitude questionnaire at roughly the time that the intervention teachers were attending their first professional development session.

The second professional development session (held several weeks into the school year) began with discussion of how the dialogic work was progressing and, where necessary, with trouble-shooting. It then moved to detailed introduction of the *Forces* module and associated resources (the other *epiSTEMe* science module was also covered during this second session). Teachers in the intervention group were given copies of the opinion questionnaire and knowledge tests, and asked to present these to their students at the designated points in the teaching cycle. Teachers in the control group received their copies by post at roughly the same time. Thereafter, contact with the teachers was via email and phone and, apart from periodic reminders, reactive to queries rather than proactive. The only exception was researcher visits to observe one *Forces* lesson in each of six intervention classes. With four *epiSTEMe* modules to be covered, resources did not permit more extensive observation, and the logistics of scheduling meant that the

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

six *Forces* classes more-or-less chose themselves. There was in other words no systematic process of selection, and in that sense the sample was random. Observations were restricted to whole-class interaction, given that it would have been difficult to monitor small-group work without undue intrusiveness.

Finally, midway through the school year, both the intervention teachers and the control group were sent copies of a background questionnaire, part of which they were asked to administer to students in the selected class while completing the other part from school records. The questionnaire covered basic demographic information, for example, student gender, social class, ethnicity and language spoken at home (since non-fluent English might compromise an intervention grounded in dialogue). Background questionnaires, attitude questionnaires (as noted also supplying background information), opinion questionnaires, and knowledge tests were returned by post after completion. Research assistants who were blind to whether instruments came from intervention or control classes coded the data in readiness for statistical analysis.

Data Preparation and Preliminary Analysis

As noted, the observational data were intended to provide information about how successfully the dialogic component had been implemented in the intervention classrooms. As the observed lessons varied in length between 42 and 54 minutes, proportionate indices were obtained, that is, the number of 6-minute periods in which each category was observed was divided by the total number of 6-minute periods across the lesson. The results are presented in Table 1, where it is clear that with a theoretical maximum of 1.00 for each cell, Classes A, B, C, and D implemented some target practices, but Classes E and F seldom did this. However, there are three practices that virtually no classes implemented.

The remaining data related to individual students. As regards opinion questionnaires and knowledge tests, full sets of data were obtained from 16 intervention teachers and 13 control teachers. The intervention teachers came from 10 different schools, and the control teachers came from 9. Two teachers implemented the intervention with two classes, and to maximize data independence only one of these classes was included in the analysis. In one case, the class was selected at random; in the other preference was given to the class that had been observed. Unfortunately, one intervention teacher and one control teacher failed to return attitude questionnaires, and two intervention teachers and one control teacher failed to return background questionnaires. To avoid undue attrition, opinion data and knowledge test data that these teachers provided were included.

The Likert scales used in the opinion and attitude questionnaires were presented using qualitative options, that is, ‘Strongly agree’, ‘Agree’, ‘Tend to agree’, ‘Neither agree nor disagree’, ‘Tend to disagree’, ‘Disagree’, and ‘Strongly disagree’. For purposes of analysis, these options were transformed to values between $+3$ and -3. Scores for negatively worded items were reversed, so that positive scores always meant favourable opinions/attitudes. Factor analysis of data derived from the

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Table 1. Proportion of observation periods in which dialogue categories occurred

	Observed classes						
	Class A	Class B	Class C	Class D	Class E	Class F	Mean (SD)
Category[a]							
Teacher asks for explanation/ clarification/reason	0.33	0.44	0.43	0.33	0.13	0.11	.30 (.14)
Student takes extended turn	0.67	0.56	0.29	0.44	0	0	.33 (.28)
Student gives reason	0.78	0.33	0.29	0.33	0	0.11	.31 (.27)
Teacher collects feedback from small-group work	0.22	0.33	0.43	0.22	0.25	0	.24 (.14)
Teacher collects >1 student view without evaluating	0.33	0.11	0.14	0.11	0	0	.12 (.12)
Teacher puts student idea/question to whole class	0.22	0	0	0	0	0	.04 (.09)
Teacher draws out difference between student ideas	0	0	0	0	0	0	0 (0)
Different perspectives discussed for >1 minute	0	0	0	0	0	0	0 (0)
Total	2.55	1.77	1.58	1.43	0.38	0.22	

[a]The ninth category (≥3 students take extended turn, give reasons, suggest new ideas, or take up another student's ideas) is omitted because of partial overlap with other categories.

opinion questionnaire indicated that responses to 14 of the 20 items were strongly loaded (.58−.82) on a single factor, which accounted for 53% of the variance. The remaining items were not strongly inter-correlated. Therefore, it was decided to base analyses on a single 14-item scale (Cronbach $\alpha = .94$) producing a score (the average item score) between −3 and +3. A single-factor structure also emerged with the attitude data, this time accounting for 45% of the variance. All 20 items loaded on this factor (one at .37 but the remainder between .57 and .82), once more warranting treatment as a single scale for purposes of analysis (Cronbach $\alpha = .93$) and producing a score (the average item score) between −3 and +3.

Most items in the knowledge tests were straightforwardly correct or incorrect. However, as coding of the explanatory statements was more subjective, two researchers independently scored the same subsample of statements (c.20%) before working separately on the remainder. Inter-judge agreement was 95% for pre-test responses, 94% for immediate post-test, and 91% for deferred post-test. Cronbach α (.67 for pre-test, .74 for immediate post-test, .77 for deferred post-test) provided reasonable grounds for regarding test scores as lying on single scales. As a result, every student was assigned scores out of 25 for each test, with two indices of learning gain computed by subtraction: (1) Pre- to immediate gain, that is, immediate post-test score less pre-test score; (2) Pre- to deferred gain, that is, deferred post-test score less pre-test score. Raw gain scores were used rather than residual gain scores because the latter are known to adjust relatively weakly for any associations between gain and pre-test.

Stronger adjustment is possible through, for example, ANCOVA on raw gains with pre-test as a co-variate (Dimitrov & Rumrill, 2003).

Results

As regards the student data, the key issue was how knowledge gain and opinions amongst the intervention group who followed the *Forces* module compared with knowledge gain and opinions amongst the control group who was taught the topic using established methods, taking account of background factors of possible relevance. Initially knowledge gain and opinions were examined as a function of background factors for each condition separately, to see whether the patterns differed across the two groups. Although the interest was in outcomes at the individual level, the students were clustered amongst 29 school classes and in principle this could affect the results. Recognizing this, the multiple regressions that constituted the within-condition analyses employed robust standard errors that were clustered on the class variable, and were conducted using *Stata Statistical Software* (Version 12, StataCorp, College Station, TX). However, when the results were checked using standard multiple regression, it became clear that the clustering was not adding precision in practice. Accordingly, the class-level variable was ignored for comparing across conditions. These between-conditions analyses employed *t*-tests and ANCOVA and were computed using the *Statistical Package for Social Sciences* Version 21 (SPSS Inc., Chicago IL).

Seven indicators of student background were deemed potentially relevant for knowledge gain with all but the seventh also potentially relevant for opinions: (1) Gender; (2) Social class; (3) Ethnicity; (4) Language used at home; (5) Pre-test score; (6) Score for the attitude questionnaire (Attitude score); (7) Score for the opinion questionnaire (Opinion score). As noted, the first four indicators and the attitude score were assessed from questionnaires, which some teachers did not return. With gender, it was usually possible to obtain the missing information from names on the knowledge tests, but there was no equivalent alternative source for the other variables. In addition, for many students information was not made available about the several indicators of social class, meaning that only one indicator was usable (crude 'eligibility for free school meals'), and even here there were substantial missing data. A broad range of categories was used to assess ethnicity, but as only a handful of students selected any specific group apart from 'white', a simple 'white' vs. 'non-white' dichotomy was used for purposes of analysis. Likewise, level of English language used at home was assessed as 'Always', 'Most of the time', 'Sometimes', 'Hardly ever', 'Never'. However, as the final three options were seldom selected, the analysis was once more based on a dichotomy: 'predominantly' (first two categories) vs. 'other' (final three categories). Descriptive data for all background characteristics are presented in Table 2.

Within-Condition Comparisons

Taking the intervention and control students separately and working with standardized scores, multiple regressions with clustered standard errors were used to examine which,

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Table 2. Student background characteristics as a function of condition

	Intervention (I)		Control (C)		
Factor	Sample[a]	Value	Sample	Value	I vs. C comparison
Gender[b]	392	46%	309	49%	$\chi^2(1) = 0.63$, ns
FSM eligibility[c]	308	16%	193	9%	$\chi^2(1) = 4.44, p = .03$
Ethnicity[d]	322	20%	250	12%	$\chi^2(1) = 5.26, p = .02$
Home language[e]	322	4%	251	2%	$\chi^2(1) = 2.36$, ns
Pre-test score[f]	398	8.46	314	10.16	$t\ (710) = 6.14, p < .001$
Attitude score[g]	336	+0.82	300	+1.15	$t\ (634) = 4.36, p < .001$
Opinion score[g]	370	+0.69	312	+0.90	$t\ (680) = 2.42, p = .02$

[a] Sample is the number of students supplying usable data for each indicator.

[b] Percentage of boys in the sample.

[c] Percentage of students eligible for free school meals.

[d] Percentage of non-white students.

[e] Percentage of students for whom English is not the main language at home.

[f] Average score out of 25 at pre-test.

[g] Average score between +3 and −3.

if any, of the seven (or six) background characteristics predicted: (1) Pre- to immediate knowledge gain; (2) Pre- to deferred knowledge gain; (3) Opinion score. Essentially, multiple regressions were conducted with all characteristics initially included, and repeated with characteristics systematically removed until 'best models' remained that contained only those predictors that were statistically significant. These models are presented in Table 3. In general, the proportion of variance that the characteristics explained was relatively low. This was especially the case for the intervention students where R^2 was 11% or 12%, indicating that responses to the *Forces* module were largely independent of student characteristics. Given the concerns summarized earlier about student subgroups, this can perhaps be regarded as reassuring. With the control students, the proportion of variance that the characteristics explained was somewhat higher, but when R^2 was between 19% and 24% can still be interpreted as modest.

Nevertheless, despite making a small contribution in absolute terms, several factors emerge in Table 3 as statistically significant predictors. With the intervention students, pre-test scores and eligibility for free school meals predicted both pre- to immediate gain and pre- to deferred gain. The negative *t*-values indicate that students with relatively high pre-test scores progressed less than students with relatively low pre-test scores, and (when 'eligible' was coded 1 and 'ineligible' was coded 0) students who were eligible for free school meals progressed less than students who were ineligible. Free school meal eligibility was also a negative predictor of opinion scores: intervention students who were eligible rated the module less favourably than intervention students who were ineligible. In addition, opinion scores were positively predicted by attitude

Table 3. Predictors of knowledge gain and opinion scores in the intervention and control samples

	t-Value	Robust standard error (non-clustered in brackets)
Intervention		
Pre- to immediate gain ($R^2 = .12$)		
Pre-test score	$-3.44, p = .005$	.08 (.06)
FSM eligibility	$-5.68, p < .001$	.12 (.15)
Pre- to deferred gain ($R^2 = .12$)		
Pre-test score	$-5.67, p < .001$	.06 (.06)
FSM eligibility	$-2.50, p = .03$	.13 (.14)
Opinion score ($R^2 = .11$)		
Attitude score	$4.55, p = .001$	.06 (.06)
FSM eligibility	$-3.39, p = .006$	.15 (.17)
Home language	$2.20, p = .05$	.24 (.41)
Control		
Pre- to immediate gain ($R^2 = .24$)		
Pre-test score	$-5.44, p < .001$	.06 (.06)
FSM eligibility	$-3.99, p = .005$	.16 (.84)
Ethnicity	$-2.53, p = .04$	.19 (.36)
Attitude score	$5.56, p = .001$	.05 (.02)
Pre- to deferred gain ($R^2 = .19$)		
Pre-test score	$-7.83, p < .001$	.05 (.07)
FSM eligibility	$-4.51, p = .003$	.13 (.23)
Attitude score	$3.35, p = .01$	.07 (.08)
Opinion score ($R^2 = .20$)		
Attitude score	$7.94, p < .001$	.06 (.05)

scores and language used at home (with 'predominantly' coded 1 and 'other' coded 0): intervention students who had relatively positive attitudes towards science and for whom English was the main language at home rated the module more favourably than other students in the intervention group. Pre-test scores and eligibility for free school meals were also relevant for the two gain scores with the control students, with both factors operating in the same direction as for the intervention group. However, attitude scores were also implicated, this time in a positive direction: control students with relatively positive attitudes to science made the most pre- to immediate gain and pre- to deferred gain. In addition, pre- to immediate gain was negatively related to ethnicity (with 'non-white' coded 1 and 'white' coded 0): in the control group, non-white students progressed less than white students. Finally, as with the intervention students, opinions scores in the control group were positively predicted by attitudes to science, although this time attitudes were the sole significant predictor.

Between-Conditions Comparisons

While the patterns that emerge in Table 3 are of intrinsic interest, they are also relevant for statistical comparison across conditions. In particular, any factors that

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

predict outcomes for both the intervention and control groups are potential confounds in between-conditions comparison *if* they are also differentially distributed across the two conditions. Factors in Table 3 that are relevant with one condition only need not be a cause for concern, nor need factors that apply with both conditions but whose distribution is equivalent. The trouble is that three factors are relevant for outcomes in both conditions: (1) Pre-test score negatively predicts pre- to immediate gain and pre- to deferred gain; (2) Eligibility for free school meals also negatively predicts both gain scores; (3) Attitude score positively predicts opinion scores. Furthermore, all three variables are differentially distributed across the conditions: as Table 2 shows, the intervention students obtained lower pre-test scores than the control students, were more likely to be eligible for free school meals, and held less favourable attitudes towards science. The implication is that pre-test score may be operating as a confound in a fashion that favours the intervention group, and eligibility for free school meals and attitudes to science may be operating as confounds in a fashion that favours the control group.

There are two possible responses to the potential confounds: one is to ignore them in the main between-conditions analyses, and the other is to incorporate them as covariates. Both strategies are unsatisfactory; the former for obvious reasons and the latter because variance that is legitimately associated with the intervention vs. control comparison will be removed from that comparison. This means that the magnitude of any between-conditions differences will be under-estimated (see Miller & Chapman, 2001).[1] Recognizing that neither strategy is above question, both were employed to compare the conditions in the hope that jointly they would provide interpretable results. Moreover, as mentioned, both were employed with standard errors no longer clustered for school class. As Table 3 shows, standard errors obtained using non-clustered data were similar to (and where different slightly larger than) those obtained after clustering, meaning that the non-clustered approach is actually more conservative than the clustered approach but nevertheless to all intents identical. Accordingly, *t*-tests were used to compare mean scores without taking account of potential confounds, and ANCOVA was used to compare mean scores with potential confounds included. The ANCOVAs relating to knowledge gain were computed first with pre-test score as covariate and then with eligibility for free school meals. They were not computed with both covariates included. Quite apart from the associations presented in Tables 2 and 3, pre-test and eligibility were themselves inter-related: students who were eligible for free school meals ($M = 8.36$, SD = 3.51) averaged significantly lower pre-test scores than students who were ineligible ($M = 9.44$, SD = 3.84), $t\,(499) = 2.18$, $p = .03$. Thus, with both variables included along with condition, variance would be inappropriately removed from multiple comparisons, rendering results uninterpretable and explaining perhaps why an exploratory attempt to encompass everything indicated possible non-homogeneity of regression slopes (Leech, Barrett, & Morgan, 2008). Tests for homogeneity were passed when the potential confounds were treated separately.

The results relating to knowledge gain are presented in Table 4. As regards pre- to immediate gain, progress was consistently higher in the intervention group than in the

Table 4. Mean knowledge gain as a function of condition

	Intervention	Control	Comparison
t-Test			
Pre- to immediate	+3.80	+2.56	t (672) = 4.46, p < .001, d = 0.35[a]
Pre- to deferred	+2.88	+1.94	t (669) = 3.30, p < .001, d = 0.26
ANCOVA (including pre-test)			
Pre- to immediate	+3.61[b]	+2.79	F(1, 671) = 8.79, p = .003, d = 0.23
Pre- to deferred	+2.66	+2.46	F(1, 662) = 0.59, ns, d = 0.06
ANCOVA (including free school meal eligibility)			
Pre- to immediate	+4.07	+2.57	F(1, 475)[c] = 21.46, p < .001, d = 0.42
Pre- to deferred	+2.74	+1.93	F(1, 469) = 6.19, p = .01, d = 0.24

[a] Effect sizes are Cohen's *d*.

[b] Means from here onwards are estimated marginal values.

[c] Reported results are based on the full samples for whom relevant data were available. Because the samples were smaller for the final two analyses, the first four were repeated using only the students for whom free school meal eligibility was known. The results are very similar, for example, d = 0.39, 0.21, 0.21, and 0.03, respectively.

control group, and the differences were statistically significant no matter which approach to analysis was followed. In both conditions, pre- to deferred gain was consistently lower than pre- to immediate gain, but remained higher in the intervention group than in the control group. However, while the differences with pre- to deferred gain were statistically significant on two of the three comparisons, the value obtained through ANCOVA with pre-test as covariate was non-significant. The results obtained from the *t*-tests relating to opinions were presented in Table 2: they indicate a statistically significant difference favouring the control group. However, this difference disappears completely in the ANCOVA with attitude score as covariate, F (1, 551) = 0.11, ns. With this ANCOVA, estimated marginal means were +0.81 in the intervention group and +0.82 in the control group.

Discussion

The sample asymmetries that compromise the between-conditions comparisons are disappointing. As noted, considerable effort was made during assignment to conditions to balance the schools that were represented in the two conditions. However, it would have intruded too far to seek to balance the Year 7 classes to which schools assigned the teachers who were included in the project, beyond the general guidance (outlined earlier) about end-of-primary-school attainment. Unfortunately, selection of classes may be where the asymmetries arose. Noting that pre-test performance was weaker in the intervention sample than in the control sample and attitudes to science less favourable, one possibility that needs to be considered is whether schools and teachers used the *epiSTEMe* intervention 'remedially' with classes which they perceived as struggling or likely to struggle. At the same time, a

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

further possibility is that the space that the control teachers needed to fit test and questionnaire administration into their normal schedule resulted in their choosing classes that were progressing through the curriculum (or expected to progress) at a relatively rapid pace. Whatever the case, it is clear that asymmetries crept in for reasons beyond the researchers' control, and the consequence is results whose meaning is partly a matter of interpretation rather than direct reading from statistical outputs.

Interpretation must of course proceed with caution, but as regards knowledge gain, there are several reasons for suspecting that the *t*-tests presented in Table 4 may reflect the true situation more accurately than any of the ANCOVAs. First, there were, as noted, two potential confounds with knowledge gain, pre-test score which operated in a fashion that could have inflated performance in the intervention group, and eligibility for free school meals whose operation could have inflated performance in the control group. The two factors operated with similar statistical power, and since they worked in opposite directions they could arguably be regarded as cancelling each other out. Second, the Chilean replication alluded to earlier (Larrain, 2013; Larrain et al., in press) produced uncannily similar results for pre- to immediate gain ($M = +3.14$ in the intervention group; $M = +2.03$ in the control group), despite this time obtaining intervention and control samples that were initially equivalent. Finally, the fact remains that, as is clear from the R^2 values in Table 3, none of the potential confounds accounted for much of the variance in knowledge gain. This latter point applies with the opinion scores too where there was only one potential confound, attitude score, which arguably biased the results towards control group positivity. Here, the most reasonable conclusion is probably that, bearing the bias in mind, the differences in Table 2 are inflated. Moreover, the key message in any event is that on average both intervention and control students held mildly positive opinions about the teaching they received on the topic of forces.

While the contrasts between the analyses presented in Table 4 are important, some aspects of the results hold good no matter which approach was followed. Encouragingly for *epiSTEMe*, there is consistent evidence throughout the table for the *Forces* module resulting in greater pre- to immediate gains than established teaching. Moreover, these gains were carried through to deferred post-test. However, there is also consistent evidence for pre- to immediate gains dissipating somewhat over the following month, so that although dissipation was also observed in the control group the intervention group's advantage at deferred post-test was less pronounced. This implies two key questions: what was responsible for the initial success and what caused the dissipation? As regards the former, it is difficult to attribute responsibility to *epiSTEMe's* general pedagogic principles. One such principle was engagement with wider experiences, and the opinion questionnaire contained items relating to this principle. Yet while responses to the questionnaire from the intervention group were mildly positive, they were no more positive than responses from the control group. A second general principle was dialogic teaching, and as recorded in Table 1, four of the observed classes implemented target practices to some extent while two were less successful. However, observations focused upon whole-class behaviour,

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

meaning that had dialogue been consequential the differences between classes would have been reflected in class-level effects upon knowledge gain and/or opinion scores. The virtual identity of the standard errors reported in Table 3 for clustered and non-clustered data argues against such effects. With the general principles unlikely to be implicated in the initial success, it is tempting to focus upon the principle that was more specific to the *Forces* module, that is, the explicit (and perhaps explicitly mathematical) emphasis on proportionality. While identification through what, in effect, is a process of elimination is clearly insufficient to establish responsibility there may nevertheless be mileage in future research that explores this principle in depth.

On the other hand, while dialogic practices may have had little relevance for the immediate gains, their characteristics (and specifically what they lacked) may have contributed to the subsequent dissipation. As is clear from Table 1, observed practices revolved around teachers asking for explanations, clarifications, or reasons and students taking extended turns and/or giving reasons, with feedback from small-group work probably containing these elements too. In other words, the students were on occasion encouraged to express ideas, and from the fifth category in Table 1 contrasting ideas were sometimes collated. However, while this means that differences were expressed (to some extent at least), the observations provide little evidence for differences being debated and eventually reconciled. No teachers drew out differences between student ideas, and in no class were different perspectives discussed for at least one minute. The implication is therefore that the gains that the intervention students displayed at immediate post-test did not result from resolution within the dialogic process, but rather depended upon events that took place independently. While it is impossible to be certain what these events involved, it is conceivable that, concerned about straying too far off-target, the teachers did in fact use the PowerPoint slides that were supplied to support understanding (see again the lower part of Figure 2) to impose an authoritative voice rather than co-ordinate this voice with dialogue. The trouble is that imposition often pre-empts the personal reflection that dialogue can stimulate (Howe, 2009a), and which allows students to make ideas their own and sometimes even progress further: growth can occur during the post-dialogue interval when differences are expressed but not extrinsically resolved (Howe, 2009b; Howe, McWilliam, & Cross, 2005).

The suggestion is therefore that dissipation occurred because pre- to immediate gains were achieved in spite of dialogue rather than because of this. Moreover while this suggestion can only be treated as speculative on the basis of present results, it receives indirect support from two additional sources. The first is the Chilean data for there, after much richer dialogue, pre- to deferred gain in the intervention group ($M = +4.71$) was actually higher than pre- to immediate gain, while pre- to deferred gain in the control group was lower ($M = +1.27$). The second source is the small-scale and informal evaluation that, as mentioned earlier, took place at the end of the development period. While the provisional nature of the data must be emphasized, it is nonetheless striking that the effect sizes obtained for learning gain in the *epiSTEMe* classes were not simply much larger than those obtained from the formal evaluation, but also more-or-less constant across the immediate and deferred

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

measures. There was in other words little dissipation. The key point is that involved over two years, the teachers who worked on the design became very familiar not just with the problem content but also with the dialogic practices. As a result, they may have come closer to emulating dialogic teaching in its ideal sense, and specifically to using the PowerPoint slides to support authoritative *reconciliation* rather than to impose an authoritative voice. Available data do not allow this possibility to be tested systematically, but as noted video-extracts were obtained from recordings made during the design stage and used for professional development during the formal evaluation. This implies that the teachers who were involved with module design did sometimes use key dialogic practices that were never observed during the evaluation.

In conclusion then, the *Forces* module resulted in higher pre- to immediate gains in knowledge and understanding than established teaching, when students were tested using instruments whose target content typically remains challenging long after the early years of secondary school. Nevertheless, while the advantages were never lost, they decreased somewhat during the period between immediate and deferred post-test. The emphasis upon proportionality, together perhaps with its mathematical representation, has emerged as the most plausible explanation of the initial success, while failure to debate ideas and resolve differences dialogically may be responsible for the subsequent dissipation. However, both possibilities are little more than speculations on the basis of the present results, so both require extensive research. Should they be endorsed, they would support the pressure from policy-makers, alluded to earlier, towards bringing science and mathematics teaching into closer alignment. At the same time, they would create new implications for policy. As noted, the two days of professional development associated with the formal evaluation are the norm for educational innovation within the UK, yet this was clearly insufficient to embed target dialogue. Intended practices may have been adopted to a greater extent during the design stage, but this stage involved eight full days of professional support and opportunities to practise over a two-year period which is far from typical. Could it be therefore that the normal situation is inadequate, and perhaps not just for *epiSTEMe* but also for many other attempts at educational reform? The implications are potentially substantial, so the issue requires careful analysis. For now, the results seem to warrant three key points: (1) there is promise in the broad approach that the *Forces* module exemplifies; (2) within that approach some (but not all) of the pedagogic principles have been pinpointed as promising; (3) in any event, more work is needed before the value can be truly appraised.

Acknowledgements

The authors wish to thank the Institute of Physics and the Economic and Social Research Council for funding, fellow members of the Targeted Initiative for feedback, James de Winter for comments on materials, Andy Tolmie and Anna Vignoles for statistical advice, Antonia Larrain for access to Chilean data, the participating students and teachers for enthusiastic engagement, and above all Stefanie Luthman, Ken

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

Ruthven, and Keith Taber who as fellow members of the *epiSTEMe* team provided unstinting support at all stages of the project.

Note

1. It is worth noting that under-estimation of between-conditions differences would also have been the consequence had residual gain scores been used rather than raw gain, even though the association between pre-test scores and gain would not then have been apparent (and therefore the interpretive challenges addressed in this paper harder to detect). There is in fact no statistical technique that avoids under-estimation when between-conditions differences are associated with variables that are also associated with outcome measures (Miller & Chapman, 2001).

References

Alexander, R. (2008). *Towards dialogic teaching: Rethinking classroom talk* (4th ed.). York: Dialogos.

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance: Classes, strategies, or rules for the balance scale task? *Cognitive Development*, *16*, 717–735.

Bransford, J., Brown, A., & Cocking, R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academies Press.

Dean, A. L., & Frankhouser, J. R. (1988). Way stations in the development of children's proportionality concepts: The stage issue revisited. *Journal of Experimental Child Psychology*, *46*, 129–149.

Department for Education. (2010). *Key stage 2 to key stage 4 (KS2-KS4): Contextual value added measure (CVA) including English and maths*. Retrieved from http://www.education.gov.uk/schools/performance/archive/schools_10/s3.shtml

Department for Education and Employment. (2001). *Framework for teaching mathematics: Years 7, 8 and 9*. London: Author.

Department for Education and Skills. (2007). *Gender and education: The evidence on pupils in England*. Retrieved from https://www.education.gov.uk/publications/eOrderingDownload/00389-2007BKT-EN.pdf

Department of Education. (1999). *Exemplary and promising mathematics programs*. Washington, DC: Author.

Dimitrov, D. M., & Rumrill, P. R. (2003). Pretest-posttest designs and measurement of change. *Work*, *20*, 159–165.

Duit, R. (2007). *Bibliography STCSE (students' and teachers' conceptions and science education)*. Retrieved from http://www.ipn.uni-kiel.de/aktuell/stcse/stcse.html

Duschl, R., Schweingruber, H., & Shouse, A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: National Academies Press.

Education Scotland. (2011). *Curriculum for excellence: Sciences (principles and practices)*. Retrieved from http://www.educationscotland.gov.uk/Images/sciences_principles_practice_tcm4-540396.pdf

Gilbert, J. (Ed.). (2006). *Science education in schools: Issues, evidence and proposals.* London: Teaching and Learning Research Programme.

Hattie, J. A. C. (2012). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.

Howe, C. (2009a). Expert support for group-work in elementary science: The role of consensus. In B. Schwarz, T. Dreyfus, & R. Hershkowitz (Eds.), *Transformation of knowledge through classroom interaction* (pp. 93–104). London: Routledge.

Howe, C. (2009b). Collaborative group work in middle childhood: Joint construction, unresolved contradiction, and the growth of knowledge. *Human Development*, *52*, 215–239.

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014

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., Nunes, T., & Bryant, P. (2010). Intensive quantities: Why they matter to developmental research. *British Journal of Developmental Psychology*, *28*, 307–329.

Howe, C., Nunes, T., Bryant, P., Bell, D., & Desli, D. (2010). Intensive quantities: Towards their recognition at primary school level. *British Journal of Educational Psychology Monograph Series*, *2*(7), 101–118.

Howe, C., Tavares Taylor, J., & Devine, A. (2012). Everyday conceptions of object fall: Explicit and tacit understanding in middle childhood. *Journal of Experimental Child Psychology*, *111*, 351–366.

Howe, C., Taylor Tavares, J., & Devine, A. (2014). Children's understanding of physical events: Explicit and tacit understanding of horizontal motion. *British Journal of Developmental Psychology*, *32*, 141–162.

Howe, C., Tolmie, A., Thurston, A., Topping, K., Christie, D., Livingston, K., … Donaldson, C. (2007). Group work in elementary science: Organizational principles for classroom teaching. *Learning and Instruction*, *17*, 549–563.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.

Larrain, A. (2013). *Argumentation in science learning in Chilean primary education: An efficacy study*. Paper presented at the 15th biennial EARLI and JURE conference, Munich, Germany.

Larrain, A., Howe, C., & Freire, P. (in press). The effect of classroom oral argumentation on middle-school science learning. Manuscript under review.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2008). *SPSS for intermediate statistics*. New York, NY: Psychology Press.

Lesh, R., Post, T., & Behr, M. (1988). Proportional reasoning. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 93–118). Reston, VA: Erlbaum.

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal*, *30*, 359–377.

Mercer, N., & Sams, C. (2006). Teaching children how to use language to solve maths problems. *Language and Education*, *20*, 507–527.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*, 40–48.

Piaget, J., Grize, J.-B., Szeminska, A., & Bang, V. (1977). *Epistemology and the psychology of functions*. Dordrecht: D. Reidel.

Piaget, J., & Inhelder, B. (1975). *Origin of the idea of chance in children*. London: Routledge and Kegan Paul.

Programme for International Student Assessment. (2009). *PISA 2009 key findings*. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa2009keyfindings.htm

Programme for International Student Assessment. (2012). *PISA 2012 results*. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm

Qualifications and Curriculum Authority. (2007). *Science: Programmes of study for key stage 3 and attainment targets*. London: Author.

Roberts, G. (2002). *SET for success: The supply of people with science, technology, engineering and mathematical skills*. London: Qualifications and Curriculum Authority.

Royal Society. (2011). *Preparing for the transfer from school and college science and mathematics education to UK STEM higher education: A 'state of the nation' report*. London: Author.

Ruthven, K., Hofmann, R., Howe, C., Luthman, S., Mercer, N., & Taber, K. (2011). The epiSTEMe pedagogical approach: Essentials, rationales and challenges. *Proceedings of the British Society for Research into Learning Mathematics*, *31*, 131–136.

Ruthven, K., Mercer, N., Taber, K., Guardia, P., Hofmann, R., Ilie, S., … Riga, F. (in press). A research-informed dialogic-teaching approach to early secondary-school mathematics and science: The pedagogical design and field trial of the *epiSTEMe* intervention. Manuscript in preparation.

Schwartz, D. L., Martin, T., & Pfaffman, J. (2005). How mathematics propels the development of physical knowledge. *Journal of Cognition and Development*, *6*, 65–88.

Schwartz, D. L., & Moore, J. L. (1998). On the role of mathematics in explaining the material world: Mental models for proportional reasoning. *Cognitive Science*, *22*, 471–516.

SCORE. (2011). *National curriculum review—call for evidence. A SCORE response to the Department for Education's call for evidence on the review of the national curriculum.* Retrieved from http://www.rsc.org/ScienceAndTechnology/Policy/EducationPolicy/SCORE-response-review-of-national-curriculum.asp

Scott, P., Mortimer, E., & Aguiar, O. (2006). The tension between authoritative and dialogic discourse: A fundamental characteristic of meaning making interactions in high school science lessons. *Science Education*, *90*, 605–631.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*(2), 1–84.

Trends in International Mathematics and Science Study. (1995). *TIMSS technical report (volume II): Implementation and analysis (primary and middle school years)*. Retrieved from http://timss.bc.edu/timss1995i/TechVol2.html

Trends in International Mathematics and Science Study. (2011). *TIMSS 2011 encyclopaedia: Education policy and curriculum in mathematics and science. Volumes 1 and 2.* Retrieved from http://timss.bc.edu/timss2011/encyclopedia-timss.html

White, R. T., & Gunstone, R. F. (2008). The conceptual change approach to the teaching of science. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 619–628). New York, NY: Routledge.

Downloaded by [University of Colorado - Health Science Library] at 13:17 13 December 2014