Research Papers in Education

Taylor & Francis Group

ISSN: 0267-1522 (Print) 1470-1146 (Online) Journal homepage: http://www.tandfonline.com/loi/rred20

# A research-informed dialogic-teaching approach to early secondary school mathematics and science: the pedagogical design and field trial of the epiSTEMe intervention

**Kenneth Ruthven, Neil Mercer, Keith S. Taber, Paula Guardia, Riikka Hofmann, Sonia Ilie, Stefanie Luthman & Fran Riga**

**To cite this article:** Kenneth Ruthven, Neil Mercer, Keith S. Taber, Paula Guardia, Riikka Hofmann, Sonia Ilie, Stefanie Luthman & Fran Riga (2016): A research-informed dialogic-teaching approach to early secondary school mathematics and science: the pedagogical design and field trial of the epiSTEMe intervention, Research Papers in Education, DOI: 10.1080/02671522.2015.1129642

**To link to this article:** http://dx.doi.org/10.1080/02671522.2015.1129642

RESEARCH PAPERS IN EDUCATION, 2016
http://dx.doi.org/10.1080/02671522.2015.1129642

Routledge
Taylor & Francis Group



# A research-informed dialogic-teaching approach to early secondary school mathematics and science: the pedagogical design and field trial of the *epiSTEMe* intervention

Kenneth Ruthven, Neil Mercer, Keith S. Taber, Paula Guardia, Riikka Hofmann, Sonia Ilie, Stefanie Luthman and Fran Riga

Faculty of Education, University of Cambridge, Cambridge, UK

**ABSTRACT**

The *Effecting Principled Improvement in STEM Education* [*epiSTEMe*] project undertook pedagogical research aimed at improving pupil engagement and learning in early secondary school physical science and mathematics. Using principles identified as effective in the research literature and drawing on a range of existing pedagogical resources, the project designed and trialled a classroom intervention, with associated professional development, in a form intended to be suited to implementation at scale. The most distinctive feature of the *epiSTEMe* pedagogical approach is its inclusion of a component of dialogic teaching. Aimed at the first year of secondary education in English schools (covering ages 11–12), the *epiSTEMe* intervention consists of a short introductory module designed to prepare classes for this dialogic teaching component, and topic modules which employ the *epiSTEMe* pedagogical approach to cover two curricular topics in each of science and mathematics. A field trial was conducted over the 2010/2011 school year in 25 volunteer schools, randomly assigned to intervention and control groups. Within the intervention group, observation of lessons indicated that the level of dialogic teaching was higher for one of the topic modules than others. Evaluation focused on the effectiveness of the topic modules, each trialled in more than 10 classes containing a total of over 300 pupils, and compared with a group of similar composition. Overall, at this first implementation, learning gains under the *epiSTEMe* intervention were no greater, although for individual topic modules the effects ranged from small negative to small positive. No difference was found between intervention and control groups either in the opinion of pupils about their classroom experience or in changes in their attitude towards subjects.

**ARTICLE HISTORY**
Received 18 March 2015
Accepted 6 December 2015

**KEYWORDS**
Dialogic teaching; pedagogical design; intervention evaluation; mathematics and science teaching; secondary school

## Introduction

Improving the quality and effectiveness of school science and mathematics education has been a prominent goal of educational policy in many countries since the international initiatives of the early 1960s (Kilpatrick 2012). Such efforts have been particularly extensive in the USA where, around the turn of the century, the National Academy of Sciences commissioned a sequence of expert panels to prepare authoritative overviews of advances in research-based knowledge about thinking, learning and teaching

**CONTACT** Kenneth Ruthven kr18@cam.ac.uk

(Bransford, Brown, and Cocking 2000; Duschl, Schweingruber, and Shouse 2007; Kilpatrick, Swafford, and Findell 2001) which had also influenced the development of *National Science Education Standards* (National Academy of Sciences 1995) and *Principles and Standards for School Mathematics* (NCTM 2000). Equally, the National Science Foundation supported development of innovative '*Standards*-based curricula' which, in the spirit of the recommendations of such syntheses, seek to help pupils explore and make sense of the material that they are learning, show that knowledge is a tool for solving problems, and foster coherent understanding of fundamental ideas and their relationships (Trafton, Reys, and Wasman 2001). Several of these programmes were judged 'exemplary' on the basis of evidence of effectiveness in multiple sites for multiple subpopulations (DoE 1999).

Against this background, an important aim of the UK Economic and Social Research Council's *Targeted Initiative on Science and Mathematics Education* was to systematise and extend the research-based knowledge available so as to inform British efforts to increase young people's levels of achievement in school science and mathematics, and to raise their rates of participation in further study and employment in these areas. As part of this initiative, the *Effecting Principled Improvement in STEM Education* [*epiSTEMe*] project undertook research-based pedagogical development aimed at improving pupil engagement and learning in early secondary school physical science and mathematics. Using principles identified as effective in the research literature and drawing on a range of existing pedagogical resources, the project designed and trialled a classroom intervention, with associated professional development, in a form intended for implementation at scale within the English educational system.

There were several reasons for *epiSTEMe*'s focus on the early secondary years. This is the phase of schooling in which pupils meet specialist teaching of mathematics and science for the first time, and it is known to be particularly important in forming young people's orientation towards further study of these subjects (Osborne, Simon, and Collins 2003). In addition, from the point of view of implementation, this phase of schooling is the earliest one in which reform becomes possible through working with relatively small cohorts of specialist teachers. Moreover, because this phase is relatively distant from the pressures of high-stakes external assessment, it offers better prospects of teachers, pupils and parents being willing to explore new approaches.

The focus of this paper is on the pedagogical design of the *epiSTEMe* intervention, and on the overall findings of a large-scale field trial of the intervention conducted in English schools.

## Key prior research on mathematics and science teaching

A range of prior research on mathematics and science teaching informed the pedagogical model developed in the *epiSTEMe* project.

### *Meta-analyses of international research on effective teaching strategies*

Just before the start of the *epiSTEMe* project, several meta-analyses had reported on the accumulated international corpus of research on effective teaching. They examined 'teaching components' in mathematics and science (Seidel and Shavelson 2007), 'teaching strategies' in science (Schroeder et al. 2007) and 'teaching programmes' in mathematics (Slavin and Lake 2008; Slavin, Lake, and Groff 2009). While these meta-analyses display important differences in their governing frameworks and specific criteria, and their results reveal significant gaps in the corpus of research available, they do provide clear indications of the relatively high effectiveness of some types of teaching component (Ruthven 2011). These types of teaching component are:

- *Domain-specific enquiry* in which classroom activity is organised around topic-related problem solving, with a focus on pupil thinking related to key concepts. This has been found to be highly effective for attainment in both subjects and for attitude in science (but is under-investigated for attitude in mathematics).

- *Cooperative group work* in which pupils work together in small groups to tackle topic-related tasks, often observing particular guidelines for interaction. This has been found to be relatively effective for attainment in both subjects and for attitude in science (but not in mathematics).
- *Enhanced context* in which teaching of a topic makes strong links to pupil experiences and interests and/or to the local setting. This has been found to be particularly effective for science attainment (but is under-investigated in other respects).
- *Active teaching* or *direct instruction* in which the teacher leads structured, interactive development of a topic, often step by step. This has been found to be relatively effective in relation to more traditional measures of attainment in mathematics (but is under-investigated in other respects).

### *Research underpinning the nationally recommended teaching approach*

In the years preceding the *epiSTEMe* project, the major influence on school mathematics and science teaching in England was the model for classroom teaching promoted by the *National Strategies* for school improvement. This 'whole-class interactive model of teaching' emphasised the importance of 'lessons hav[ing] clear objectives and [being] suitably paced' and of 'a high proportion of each lesson [being] spent on direct teaching' (DfEE 2001, 6 & 26). Support for this model, corresponding to the teaching component of active teaching or direct instruction identified by the meta-analyses, came from an earlier research synthesis which highlighted an American tradition of process-product research on effective mathematics teaching (Good, Grouws, and Ebmeier 1983). This was claimed to accord both with a much smaller body of British research, and with the judgement of English school inspectors in their contemporary reports on the school system (DfEE 1998; Reynolds and Muijs 1999).

However, recognising that this model had been validated primarily in relation to the teaching of basic skills, the advocates of whole-class interactive teaching acknowledged the relevance of more recent research which indicated that 'additional classroom processes may be needed to enhance higher order thinking', listing these as 'a focus on meaning and understanding … direct teaching of higher level cognitive strategies and problem-solving, and co-operative small group work' (Reynolds and Muijs 1999, 281). Although not using identical terminology, these suggestions accord with the conclusions drawn from the later meta-analyses about the proven effectiveness of teaching components of domain-specific enquiry and cooperative group work.

### *Systematic national reviews of research on user-prioritised issues in teaching*

Prior to the *epiSTEMe* project, the *Evidence for Policy and Practice Initiative* (Bennett et al. 2005) had conducted several systematic reviews of British research focusing on specific issues of mathematics and science teaching that a range of users, notably practitioners and policymakers, judged particularly worthy of attention. Prominent amongst these were the role of classroom dialogue and discussion in supporting teaching approaches with components corresponding to domain-specific enquiry and cooperative group work.

Kyriacou and Issitt (2008) investigated what characteristics of teacher-initiated teacher–pupil dialogue made it effective in promoting conceptual understanding in mathematics at upper-primary and lower-secondary school levels. They concluded that for such dialogue to be effective in supporting approaches of the domain-specific enquiry type, it needs to go beyond traditional classroom interaction in an initiation–response–feedback pattern, to display features such as 'focusing attention on mathematics rather than performativity; working collaboratively with pupils; transformative listening; scaffolding; enhancing pupils' self-knowledge of how to make use of teacher-pupil dialogue as a learning experience; [and] encouraging high quality pupil dialogue' (Kyriacou and Issitt 2008, 13).

Bennett et al. (2010) synthesised the findings from several earlier systematic reviews which focused on different aspects of the use of small group discussion in secondary school science teaching, thus examining an important aspect of cooperative group work. They reported that 'groups function more

purposefully, and understanding improves most, when specifically constituted such that differing views are represented, when some form of training is provided for pupils on effective group work, and when help in structuring discussions is provided' (69). They concluded that 'for small group discussions to be effective, teachers and students need to be given explicit teaching in the skills associated with the development of arguments and the characteristics associated with effective group discussions' (69).

### *Underpinning research on classroom dialogue and dialogic teaching*

A key body of research underpinning the findings of these systematic reviews pointed to the value of dialogic small-group and whole-class discussion in encouraging pupils to talk in an exploratory way and to consider different points of view (Howe and Tolmie 2003; Howe et al. 2007; Mercer and Sams 2006; Mercer et al. 2004). In particular, findings from a long-standing programme which developed a discourse-based approach that teachers used successfully to promote 'thinking together' in science (Mercer et al. 2004) and mathematics (Mercer and Sams 2006) indicated that pupils could be enabled to use talk more effectively as a tool for reasoning, and that talk-based group activities could help develop individuals' mathematical and scientific reasoning, understanding and problem-solving. The conception of teaching-and-learning that underpins such an approach is essentially a Vygotskian, sociocultural one, whereby students are inducted into the communities of discourse of science and mathematics, and in which dialogue can play an important role in enabling conceptual change (Mercer and Littleton 2007; Scott, Asoko, and Leach 2007). The key mechanisms have been identified as *teacher-led interaction with pupils* which plays a crucial role in inducting pupils into the discourses associated with particular knowledge domains while *peer group interaction between pupils* provides more 'symmetrical' opportunities for pupils to examine existing understandings and make changes where necessary, and to relate their developing understanding to their everyday world. Such developments have contributed to the emergence of the notion of *dialogic teaching.*

Indeed, there has been a long-standing interest in science education in the way teachers use talk in classrooms to support the initiation and development of learners' concepts about aspects of the natural world and in particular those many scientific concepts which do not have concrete observable referents (Lemke 1990; Ogborn et al. 1996). The widespread adoption of a broadly constructivist perspective on student learning and the recognition that students commonly develop alternative conceptions of scientific ideas has led to a focus on the need for teaching to be based on dialogic interactions that allow teachers to monitor, and seek to modify, aspects of students' developing conceptualisations (Scott 1998). Dialogic talk has therefore come to be seen as core feature of science teaching that incorporates formative assessment, that is ongoing teacher assessment of thinking intended to support intended learning during teaching (Black and Atkin 2014).

Scott, Mortimer, and Aguiar (2006) point out that any form of multispeaker classroom discourse can be regarded as dialogic, in a weaker sense, inasmuch as utterances take account of previous contributions and anticipate the responses of others. However, they also identify a stronger sense in which classroom discourse becomes dialogic only when speakers give more explicit recognition to different points of view and seek to compare these. This leads to a distinction between two dimensions of classroom discourse. The first dimension, corresponding to the weaker sense, is that of the *interactive–non-interactive* quality of the pattern of talk. The second dimension, corresponding to the stronger sense, is that of the *authoritative-dialogic* framing of the substance of talk. Scott, Mortimer and Aguiar acknowledge that there is an important place for authoritative discourse in science and mathematics teaching as part of the process by which pupils are inducted into specialised and complex modes of thinking. Equally, however, they argue that there is as important a place for dialogic discourse, used here in the stronger sense, which encourages the expression of ideas by pupils, identifies contrasting points of view, and engages seriously in examination and comparison of the reasoning associated with these. They suggest that dialogic teaching is likely to shift in its focus over the course of a sequence of lessons: whereas at the start of a lesson sequence, the teacher might be eliciting pupils' existing

thinking about a topic, by the end of the sequence, the teacher is more likely to be encouraging pupils to discuss how to apply a newly learned idea in a novel context.

## The *epiSTEMe* intervention

The focus of the *epiSTEMe* project was on designing and trialling a classroom intervention, with associated professional development, in a form intended to be suited to implementation at scale.

### *The pedagogical model*

While the *epiSTEMe* pedagogical model places a strong emphasis on exploratory dialogic talk, both in small groups and whole class, it also makes provision for codification and consolidation of key ideas to take place later (Ruthven et al. 2012). In the exploratory phase, domain-specific enquiry is employed to support informal development of target concepts. Dialogic small-group and whole-class discussion provides opportunity for pupils to express their thinking about a problem situation and to examine different perspectives on it. During such discussion, the teacher's principal role is to support the dialogic quality of contributions by pupils and exchanges between them. For codification, the teacher's role becomes a more authoritative one of explaining accepted mathematico-scientific approaches to the problem situation through active teaching which, although it may be interactive and take account of the thinking displayed during the earlier exploration phase, accords the accepted approach a privileged status. Finally, for consolidation, pupils tackle related problem situations more independently, with the teacher's role becoming one of checking pupil understanding and providing developmental feedback.

Following several of those US *Standards*-based programmes that have been judged exemplary, the domain-specific enquiry employed in *epiSTEMe* lessons is organised around carefully crafted problem situations. Such problems are devised with a view to developing key disciplinary concepts, and (in view of the promising findings noted earlier for the teaching strategy of enhanced context) are often posed in ways which seek to appeal to widely shared pupil experiences and interests. The staging of the problem also aims to take account of what is known about informal knowledge and thinking related to a topic. To achieve this, the teams which devised each of the *epiSTEMe* lesson sequences drew extensively on research on the epistemology of the topic concerned and on didactical approaches to it, as well as on research on conceptual development in the topic and common forms of fallacious reasoning about it. These resources informed the treatment of the topic throughout the lesson-sequences.

Thus, what makes the *epiSTEMe* approach pedagogically distinctive is its blending of the teaching components of domain-specific enquiry, cooperative group work and enhanced context, guided by the overarching notion of dialogic teaching. This approach was intended to provide a stronger basis for the more interactive and adaptive components of teaching which, on the evidence of inspection surveys, were underdeveloped in English professional practice (OfStEd 2008a, 2008b).

### *The intervention apparatus*

The project devised the apparatus of the *epiSTEMe* intervention to support teachers and departments in developing teaching along the lines of the *epiSTEMe* pedagogical model, without requiring significant reorganisation of work and substantial investment of time on their part. The project's orientation was towards what might be termed 're-design' research that recognises that design for implementation at scale needs to take account of the existing state of the system: notably the people, structures, resources and practices already in place. The classroom intervention was relatively modest in scale and scope, and packaged as a viable substitute for modules currently widely taught in schools during Year 7, the first year of secondary education in England (during which pupils reach the age of 12 years).

The development and refinement of the *epiSTEMe* intervention involved working with science and mathematics teachers from five partner schools to devise and pilot classroom activities reflecting the

*epiSTEMe* pedagogical model. Lesson observation and close interaction with participating teachers provided valuable informal feedback and evidence. Examples and insights gained from these sources helped not only in refining the lesson sequences and supporting materials, but also in devising professional development activities. In particular, to better assist teachers to translate the notion of dialogic activity into practical action, we devised ways of incorporating discussion prompts and supports into classroom materials, and undertook analysis to identify examples and strategies which would enable us to communicate a more powerful operational model of the dialogic teaching aspect of the intervention (Ruthven, Hofmann, and Mercer 2011).

The *epiSTEMe* apparatus consists of the following elements. The introductory module is designed to build teacher and pupil understanding of the value of talk and dialogue in supporting subject thinking and learning, and to develop rules and processes that support effective small-group and whole-class discussion. This addresses the crucial need, identified in the systematic reviews discussed earlier, to cultivate, amongst both teachers and pupils, productive shared norms of participation in small-group and whole-class discussion, and the capacity to make use of dialogue to promote effective learning. Two further topic modules in each subject are designed to support and capitalise on such use of talk and dialogue, and to instantiate the full *epiSTEMe* pedagogical model. The two *epiSTEMe* topic modules in science focus on Electrical Circuits (Taber et al. 2015) and Forces and Proportional Relations (Howe et al. 2015a); those in mathematics on Fractions, Ratios and Proportions (Howe et al. 2015b), (subsequently referred to as Ratios), and Probability (Ruthven and Hofmann 2013). While there is not space to cover the design rationale and eventual form of each of these modules in this paper, these are discussed fully in the cited publications, and the teaching notes for each module are available online at http://www.educ.cam.ac.uk/research/projects/episteme/materials.html.

To take realistic account of the conditions under which innovation typically takes place in the English educational system, the duration of the professional development supporting the intervention was restricted to two days to reflect the limited release time usually available to teachers in English schools in preparing to implement such an initiative. While it is known that intensive and sustained support is needed for change in teaching practices (Supovitz and Turner 2000), all education systems are obliged to find ways of providing professional development within limited time and resources. The *epiSTEMe* intervention was designed to incorporate a number of features known to support professional learning under such circumstances. First, the module teaching materials are designed to be educative in the sense of supporting teacher development as well as classroom activity (Davis and Krajcik 2005). Second, the sequence of two one-day professional development events is designed to make links with intermediate activity carried out by teachers in school with their own classes. The first event focuses on dialogic teaching and on how the introductory module supports its development. Then, after teachers have undertaken the introductory module with one of their classes, the second event debriefs this experience and examines how the topic modules in their subject incorporate the pedagogical principles and processes of the *epiSTEMe* model. Third, in recruiting schools for the field trial, the project team emphasised the value of several teachers participating, particularly pairs within a subject, to provide a stronger basis for the collaboration between teachers within a school which is known to support implementation of innovations (Louis, Marks, and Kruse 1996).

## An illustrative dialogic activity

To illustrate the dialogic dimension of the *epiSTEMe* pedagogical approach more concretely, we will use the example of a lesson activity that examines a simple probabilistic model of genetic inheritance. The key genetic ideas underpinning the model (as shown in the first slide in Figure 1) are introduced to the class in an interactive style. Pupils are often surprised to learn of the two earlobe types; typically they show great interest in knowing which type they and their classmates have! The questions on the slide are designed to support collective extraction and organisation of information from the text: one that incorporates features of mathematico-scientific language that pupils need to learn about but which many find challenging at this stage. The probabilistic aspect of the model is then introduced

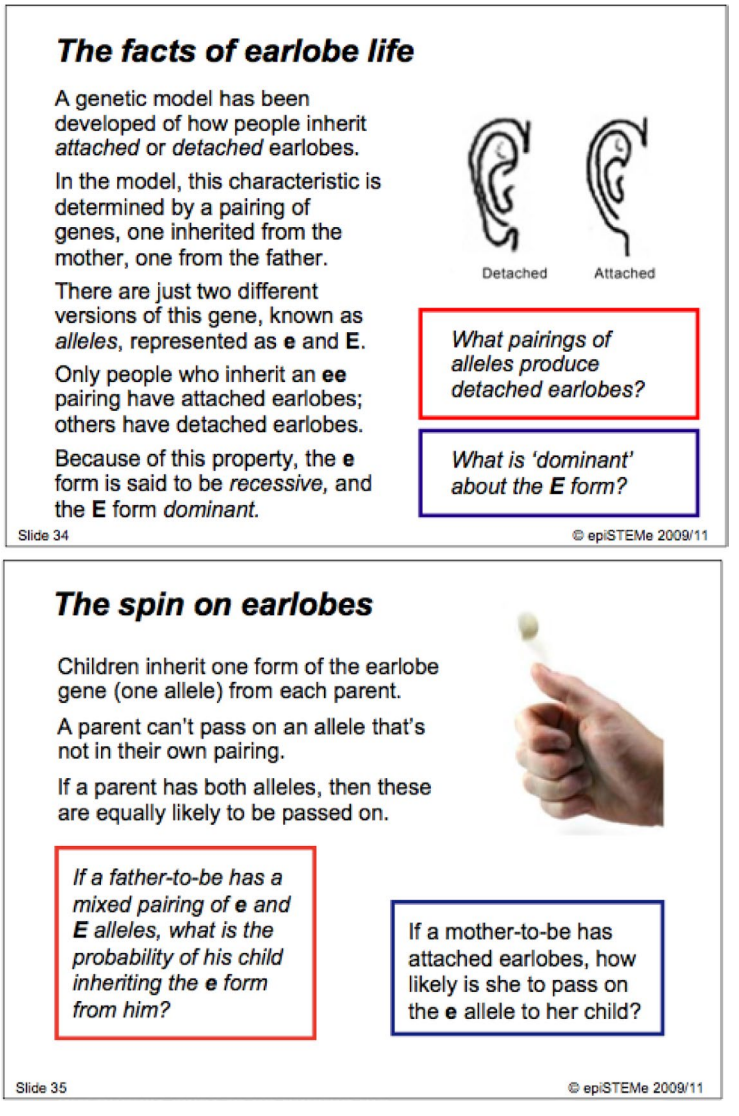


**Figure 1.** The earlobe lottery.

(as shown in the second slide). Over the course of this introduction, the teacher aims for pupils to come to grasp, first the distinction, and second the relation, between allele pattern and earlobe type. It is also not unusual for some pupil to pose the question of whether attached earlobes will eventually die out; this is acknowledged to be an interesting question that it may be possible to address in due course. Typically, too, some pupil asks whether both problems on the second slide concern the same child; this provides a good lead into the problem that pupils are then asked to work on, initially in small groups: *A couple are expecting their first baby. Both parents have a mixed pairing of e and E alleles. How likely is their baby to have this same pairing?*

An important ground-rule for small-group discussion is that pupils should try to come to an agreed position; even if they are unable to achieve this goal, honouring it calls for them to engage with points of view other than their own, and to develop an argument in support of their position. Once most groups have formulated some kind of response, the activity switches to a whole-class plenary in which their varying answers and arguments are elicited. Typically, there is a patent need for further

whole-class discussion, because different groups have arrived at what are clearly contrasting answers. Moreover, each answer arises from a distinctive pattern of reasoning: an everyday model of inheritance in which 'children take after their parents' (leading to an answer of 100%) as well as variant patterns of probabilistic reasoning about the outcome space under the scientific model of genetically mediated inheritance (leading to answers of 1/3 and 50%). These three responses represent, respectively, the predominant everyday misconception about inheritance of characteristics, the predominant lay misconception about outcomes in a simple repeated trial, and finally the accepted and coordinated mathematico-scientific conceptions.

The first *epiSTEMe* professional development event employs the videotaped example of a plenary review of this problem to examine how teachers can support quality of classroom discussion. Analysis of this example helped to concretise teaching strategies and tactics that can assist and develop dialogic exchange (Ruthven, Hofmann, and Mercer 2011). This analysis informed our choice of a sequence of short video episodes to stimulate discussion with and between teachers, with the classroom dialogue transcribed to encourage attention to the fine grain of pupils' thinking and of the teacher's participation in exchanges.[1] The emphasis is on encouraging teachers to 'read' what is taking place as each episode unfolds so as to understand how pupils are responding and reasoning, to analyse the quality of dialogic interaction, and to anticipate accordingly how the teacher might productively shape events and ideas. While the research analysis informs the contributions of members of the *epiSTEMe* team to the discussion with teachers during professional development, that analysis is not explicitly presented to them.

Supporting dialogic exchange is the aspect of the *epiSTEMe* pedagogical model that teachers reported finding particularly challenging. Because this approach emphasises developing mathematico-scientific reasoning as its goal, not simply securing task performance, it requires significant shifts beyond the received ideas and habitual reflexes of established practice. For example, a dialogic approach calls for the teacher to be prepared to give time to multiple pupil contributions including some which can be persuasively fallacious or poorly formulated. To sustain productive discussion, the teacher must be able to identify and interanimate the thinking behind different pupil responses, and steer progression in reasoning without closing down discussion.

## Research design

The field trial was designed as an experimental study, randomised at the school level between intervention and control groups consisting of intact classes. Judgements about effectiveness of the intervention were based on learning gain by pupils (inferred from topic proficiency tests administered to each class before and after undertaking a topic module), immediate pupil reaction (inferred from module opinion questionnaires administered to each class after completion of a topic module), and longer term dispositional change in pupils (inferred from subject attitude questionnaires administered to each class close to the start and end of the school year). In addition, evidence about implementation of the intervention and potential confounding factors was gathered through classroom observation and teacher questionnaires. Each of these aspects will now be described in greater detail.

### *Sample and implementation*

The field trial was conducted over the 2010/2011 school year. The intention was to recruit 30 schools to participate, together providing 60 teachers/classes in each subject, so as to yield a structured sample of sufficient size to afford a hierarchical analysis of adequate statistical power. Recruitment of schools took place from March onwards in the prior school year, with a view to ensuring good time for planning and preparation before the start of the new school year in September. In practice, it proved necessary to continue recruitment right up to the end of the prior school year in July, and to compromise on the size and structure of the sample. In particular, while the originally stipulation was that schools should nominate two science teachers and two mathematics teachers, it became clear that insisting on this

**Table 1.** Number of schools and teachers participating in each subject, by condition.

| Topic | Intervention | | Control | |
|---|---|---|---|---|
| | Schools | Teachers | Schools | Teachers |
| Science | 10 | 16 | 11 | 20 |
| Mathematics | 10 | 15 | 10 | 19 |

would result in far too few schools participating in the trial. Consequently, both the two-subject and teacher-pair requirements were relaxed.

An open invitation was sent to schools across the Eastern region of England and into North London to attend a briefing session about the project.[2] This session provided a broad overview of what participation in the project would involve. In particular, only half the participating schools were to implement the *epiSTEMe* intervention during the following year, i.e. constitute the intervention group. The remainder would be asked to act as a control group, which would involve teaching via established methods, while administering the *epiSTEMe* evaluation instruments. On completion of the field trial, control schools would be offered the same *epiSTEMe* materials and professional development provided to the intervention group.

The decision to randomise at the school level, and so to assign all teachers from any single school to the same condition, was made for two principal reasons: first, to mitigate against potential problems of intervention 'leakage' where the same school also hosts teachers belonging to the control group (Plewis and Hurry 1998); second, to provide a stronger basis for the collaboration between teachers within a school. However, one effect of the relaxation of the teacher-pair requirement within a subject was that around half of participating schools nominated only a single teacher in any subject, thus losing these collaborative opportunities.

All schools subsequently completing the application process were assigned to an experimental group using an approach in which they were paired according to school type and contextual value-added score (based on a standardised index of teaching efficacy (CVA2-4) used nationwide in England (Department for Education 2010)), and then randomly allocated to the intervention or control group.[3] One school withdrew prior to the start of the field trial because of staffing shortages. This yielded 25 participating schools: 12 in the intervention group and 13 in the control. Thus, while the number of schools participating came close to original intentions (25 rather than 30), as a result of the relaxation of participation requirements noted above the number of teachers/classes fell well short (34 in mathematics, 36 in science, rather than 60 in each) as shown by subject in Table 1.

Schools were asked to nominate teachers who would be teaching a first-year secondary class (Year 7 in England). Assignment of teachers to classes took place vicariously within each school without any involvement of the research team; where, occasionally, nominated teachers were assigned two such classes, only one was retained for this study. No specific guidance was given to schools over the choice of teachers and classes, apart from recommending that only a minority of pupils in any participating class should have performed well below national expectations in the subject at the end of primary school (i.e. below Level 4 in national assessment, attained by around 80% of pupils in England). This recommendation was made to ensure that participating classes would contain sufficient pupils able to confidently read the *epiSTEMe* materials and engage with the *epiSTEMe* tasks, supporting others in their class and group if necessary.

In order to make arrangements for teachers in the intervention group to start the professional development activity during the end-of-school-year period when schools often find it easier to release them, as well as making it possible for them to then undertake planning for the coming school year over the vacation period, it was necessary to assign schools to a condition sufficiently early. Thus, in early June, the 14 schools which had already confirmed participation were assigned to a condition, with further schools who confirmed later being assigned in subsequent batches until the end of July. However, in many schools, decisions about the formation of classes and the allocation of participating teachers to

them had not been made at the point of assignment, with some schools delaying these until close to the start of the school year, even occasionally changing the participating teachers because of staffing pressures. While the project team could provide supplementary professional development sessions to accommodate this situation, nothing could be done to address a potential threat to validity which arose from decisions about participating teachers and classes being modified, or established for the first time, by schools after they had been randomly assigned to a condition: namely the risk of bias being introduced through decision-making at school level being influenced by this assignment. This pointed to the importance of establishing, as a preliminary to later analysis, whether the treatment groups were indeed equivalent through examining the initial standing of the participating classes, and of then taking appropriate mitigating action if not.

The protocol for the field trial stipulated that teaching of the target topics in both groups of schools should take place at the appropriate point in the school's existing scheme of work for the year. Intervention schools would follow the *epiSTEMe* module, control schools the existing module or modules on that topic in their scheme of work. In the rare case of there being more than one such module, the evaluation instruments would be administered in relation to the most substantial one relevant to the learning objectives for the topic. These stipulations were intended to minimise disruption to the established sequence of topics taught in each school, and to secure a sound comparison between 'business as usual' in the control group and first implementation of each *epiSTEMe* module in the intervention group. The function of the control group was to provide a baseline of existing practice. While this practice might vary between schools and teachers (and the resources available to the project did not permit this to be investigated), a key intention behind randomisation was to produce intervention and control groups whose profiles of existing practice could be taken to be equivalent (even if not uniform).

There was a degree of attrition of classes, averaging 17% in the intervention group and 31% in the control group, as shown by topic module in Table 2. In some cases, participating teachers moved school, or were reassigned from their Year 7 class because of staffing shortages affecting older examination classes. In other cases, teachers forgot to administer instruments, or simply did not teach a particular topic.[4] Once account is taken of cases in which no returns were made at all, the classes and teachers participating within a subject were substantially the same for each topic. Table 3 summarises, for each topic, the number of classes and pupils from which a portfolio of completed instruments was received. All such classes were included in the subsequent analyses. In particular, sample sizes were sufficiently large that any difference between intervention and control groups which approached what is conventionally regarded as a small effect size (i.e. 0.2) would be statistically significant at a conventional level (i.e. 5%).

## Instruments and measures

As no suitable standard instrumentation was available for use across participating schools, it was necessary to develop suitable instruments, drawing on whatever useful precedents were available, and taking account of the particular circumstances of this study. All instruments were extensively trialled and refined during the development stage of the project.

### Topic proficiency tests

The topic proficiency tests contained a mixture of multiple-choice and open-response items, and were designed to be capable of being completed by pupils in between 10 and 15 min. Items were directed at aspects of the topic specified in national guidance on curriculum and assessment as learning objectives for pupils at this level. During the course of the teaching of a topic, tests were administered on three occasions: a pre-test during the first lesson to assess initial proficiency; an immediate post-test during the final lesson to assess proficiency upon completion of teaching; and a deferred post-test, unannounced about one month after the immediate post-test, to assess retained proficiency. The tests and test items had been developed and piloted in the earlier stages of the project. Items were extracted

**Table 2.** Attrition rate of teachers/classes for each topic, by condition.

| | Intervention | | Control | |
|---|---|---|---|---|
| Topic | Rate | as % | Rate | as % |
| Electricity | 2/16 | 13 | 8/20 | 40 |
| Forces | 0/16 | 0 | 7/20 | 35 |
| Probability | 4/15 | 27 | 6/19 | 32 |
| Ratios | 4/15 | 27 | 3/19 | 16 |

**Table 3.** Number of classes and pupils participating for each topic, by condition.

| | Intervention | | Control | |
|---|---|---|---|---|
| Topic | Classes | Pupils | Classes | Pupils |
| Electricity | 14 | 369 | 12 | 313 |
| Forces | 16 | 419 | 13 | 335 |
| Probability | 11 | 308 | 13 | 376 |
| Ratios | 11 | 311 | 16 | 463 |

and sometimes adapted from relevant research studies and published national tests.[5] In Electricity and Probability, each version of the tests had the same structure, with identical or parallel items piloted before use. Although the Forces and Ratios tests were not structured in this way, evidence from piloting had been used to generate statistically equivalent versions. Coefficients of internal consistency (Cronbach alpha lying between 0.76 and 0.87) indicate that these tests provided acceptably reliable indices of pupil subject proficiency.

### *Module opinion questionnaire*

The same module opinion questionnaire was used for each topic, customised simply by inserting the name of the relevant topic in the heading of the questionnaire. Administered in the final lesson, prior to the immediate post-test, the questionnaire consisted of statements to which response was invited on a seven-point Likert scale ranging from 'Strongly agree' to 'Strongly disagree'. Scores for negatively worded items were reversed, so that positive scores always denoted favourable opinions. Factor analysis indicated that, across all topics, responses to 14 items were strongly loaded (.61 to .79, averaging .73) on a single factor, which accounted for 53% of the variance. This allows the formation of a single 14-item opinion measure. These items consisted of seven component pairs concerned with interest experienced in the topic and future intentions towards it, improvement in understanding of the topic and in capacity to explain it, influence on valuation of the topic and appreciation of wider application, and stimulus to thinking provided by work on the topic. Coefficients of internal consistency (Cronbach alpha between 0.93 and 0.94) testify that these questionnaire-based measures provided highly reliable indices of pupil opinion.

### *Subject attitude questionnaire*

A common subject attitude questionnaire was used for both subjects, customised by inserting the name of the relevant subject into questionnaire items. Administered twice, towards the start and end of the school year, the questionnaire used the same seven-point Likert scale as did the opinion questionnaire. Likewise, scores for negatively worded items were reversed, so that positive scores always denoted favourable attitudes. Factor analysis indicated that responses to 20 items were strongly loaded on a single factor which accounted for between 45 and 53% of the variance on the four occasions of administration (with average item loading ranging from .66 to .72 over these occasions). This allows the formation of a single 20-item attitude measure. The four components, each of five items, are concerned with respondents' ratings of personal ability in the subject, personal enthusiasm for the subject, prospective involvement in the subject, and value outside study of the subject. Coefficients of

internal consistency (Cronbach alpha between 0.93 and 0.95) testify that these questionnaire-based measures provided highly reliable indices of pupil attitude.

### *Pupil background data*

Basic demographic data about participating pupils was gathered through a pupil questionnaire and via class teachers from school records. Key pupil background variables were those of gender, free-school-meal status (the crude, but available, proxy for social class), ethnicity (reduced to a dichotomy according to whether pupils identified themselves as White or not), and English-language status (reduced to a dichotomy according to whether pupils reported using English at home 'Always' or 'Most of the time' as against 'Sometimes', 'Hardly ever' or 'Never'). Returns of this data were lower and less complete than others, leading to reduced sample size when such variables were included in analyses. Fortunately, preliminary analyses indicated that these variables did not give rise to any potential confounds that the main analyses ought to take account of.

### *Dialogic teaching observation*

A classroom observation instrument was used to collect evidence of the extent to which classroom interaction in lessons displayed characteristic features of dialogic teaching. The research protocol specified 24 lesson observations, spread evenly across the four modules, with no class or teacher being observed more than once. Thus, 24 of the 28 intervention classes which completed at least one topic module were observed. Module designers indicated which lessons were particularly expected to feature dialogic activity, and the observer ensured that it was one of these lessons which was observed in each case. A balance was maintained for each topic between teachers teaching it as their first *epiSTEMe* module and as their second. Otherwise the choice of lessons was fortuitous according to the timing of teaching and the availability of the trained observer. While teachers knew that the research team was interested in observing one of a selected list of lessons, they were not aware of the particular focus of the observation on the dialogic markers. Unfortunately, project resources did not permit any observation of lessons in the control group: in retrospect, this was clearly a weakness of the research design.

The observational instrument focused on markers of classroom dialogic activity, specified as shown in Table 4. The first two markers are concerned with teacher solicitation of ideas from pupils (TSolC, TSolF), the next two with articulation of ideas by pupils (PArtR, PArtE). A further two markers are then concerned with multiplicity of pupil ideas (PMul, TMul), and another with teacher spotlighting of some pupil idea (TSpot). The final two markers are concerned with the comparison of ideas (PCom, TCom). These markers were chosen from a much longer list on the basis of satisfactory levels of inter-observer agreement during analysis of video-recorded lesson sequences from the development stage of the project.[6]

The observational procedure was only applied to whole-class interactions because it was not realistic (given classroom settings where physical layout and noise levels would require an observer to sit with one target group) to apply it to small-group interactions without selecting one particular group and impinging on their activity, which we judged likely to compromise validity. Thus, the observation process was applied to 4-min units of whole-class activity (with the subsequent 2 min used for coding). Essentially, every observational unit was coded for the presence (or absence) of each marker. During the field trial, all classroom observations were carried out by the researcher responsible for developing and refining the instrument, providing some assurance about consistency of coding.

### *Opportunity to learn indicators*

Although the design of the evaluative instruments and the protocol for the field trial sought to anticipate and pre-empt potential confounding factors, evidence was gathered through post-module teacher questionnaire[7] about two specific issues known to be potential sources of bias in this type of experimental evaluation, both potentially leading to pupils having differential opportunity to learn: the cumulative duration of the topic lessons (Scheerens et al. 2013) and the match of tests to coverage of the topic by the class (Slavin and Lake 2008). In particular, because control schools followed their own

**Table 4.** Observational markers of dialogic talk.

| Code | Marker of dialogic talk |
|---|---|
| TSolC | Teacher asks for explanation/clarification/reason |
| TSolF | Teacher collects feedback from planned small group work for at least 1 min |
| PArtR | Pupil gives a reason |
| PArtE | Pupil takes an extended turn |
| PMul | Number of pupils who do any of these things is three or more:<br>Takes an extended turn; Gives reason; Suggests a new idea/response to a task; Takes up another pupil's idea |
| TMul | Teacher collects at least two pupil views without evaluating them |
| TSpot | Teacher puts a pupil idea/question to whole class to listen or respond to |
| PCom | Different perspectives are discussed for at least 1 min |
| TCom | Teacher draws out difference between pupil ideas |

**Table 5.** Teacher reports of total lesson time* spent on topic by their class, by condition.

| Topic | Intervention | | Control | | Difference |
|---|---|---|---|---|---|
| | Mean | St. dev. | Mean | St. dev. | Stat. sig.† |
| Electricity | 753 | 169 | 595 | 202 | $p = .017$ |
| Forces | 762 | 172 | 553 | 163 | $p = .003$ |
| Probability | 451 | 35 | 309 | 144 | $p = .024$ |
| Ratios | 404 | 95 | 504 | 143 | *ns* |

*In minutes.
†Undirected *t*-test.

schemes of work for Year 7 (relating to an official programme of study covering the three-year period from Year 7 to Year 9), it was possible that they would devote different amounts of time to a topic or emphasise different aspects. And, because the topic tests had been compiled by the module designers, they might inadvertently prove closer to the emphases and approach of the intervention module.

Accordingly, a post-module teacher questionnaire asked how much classroom time in total had been spent on the lessons on the topic (Table 5). There proved to be considerable variation between classes within groups, other than for the Probability intervention group. Moreover, for the three topics of Electricity, Forces and Probability, the mean time spent on the topic by classes in the intervention group was substantially (and statistically significantly) greater than that spent by the control group. The only topic for which time proved to be equivalent between the two experimental groups was Ratios.

Similarly, the teacher questionnaire asked teachers to rate each test item according to 'how suitable [it] is for this class given its experience of the topic this school year' (Table 6). Mean teacher ratings of the test items for the three topics of Electricity, Ratios and Probability proved to be equivalent for intervention and control classes. For Forces, however, the difference was a statistically significant one of around three quarters of an interval on the six-interval Likert scale, suggesting that teachers perceived the test as better matched to the experience of the intervention group, a perception validated by further investigation of the closeness of match.

Ratios, then, was the only topic for which intervention and control conditions were equivalent in both these respects. For Forces, on the contrary, the intervention condition was favoured in both respects, and strongly so in terms of test match. For this reason, Forces has been excluded from the further analyses reported in this paper. Because data about mean time spent on the topic were missing from around 20% of returns, this was not included as a variable in the main analyses; instead, the findings reported here need to be borne in mind when the results of those analyses are interpreted.

## *Approach to statistical analysis*

Before comparing the experimental conditions, preliminary analyses of criterion measures were carried out within each condition. Following Dimitrov and Rumrill (2003), analyses were based on gain

**Table 6.** Teacher ratings* of suitability of test items given coverage of topic by their class, by condition.

| Topic | Intervention | | Control | | Difference |
|---|---|---|---|---|---|
| | Mean | St. dev. | Mean | St. dev. | Stat. sig.† |
| Electricity | +2.17 | 0.47 | +1.57 | 0.67 | *ns* |
| Forces | +1.96 | 0.86 | +1.22 | 0.77 | $p = .026$ |
| Probability | +1.51 | 0.92 | +1.71 | 0.79 | *ns* |
| Ratios | +1.80 | 0.66 | +1.45 | 0.75 | *ns* |

*On a scale from −3 [Strong disagreement] to +3 [Strong agreement].
†Undirected *t*-test.

scores where available. The compromises necessary to secure a reasonably large sample, followed by the attrition from it, meant that the scale and structure of the final data-set made a hierarchical approach to analysis questionable. Nevertheless, although the approach taken to analysis focused on outcomes at the individual level, the hierarchical structure of the data-set was recognised, with pupils nested within classes, often one class to a school. Hence, the multiple regressions on which these initial within-condition analyses were based employed robust standard errors clustered on the class variable (using the cluster sandwich estimator), and were conducted using *Stata Statistical Software*. However, when results were checked using standard multiple regression, it became clear that the clustering was not adding precision in practice: in particular, standard errors obtained using non-clustered data were similar to those obtained after clustering. Accordingly, the class-level variable was ignored for comparing across conditions.

The within-condition analyses found that pre-test score was a strong predictor of learning gain for all topics, with a negative correlation between them. Similarly pre-quest attitude was found to be a strong predictor of attitude change in both subjects, again with a negative correlation. This is a widely observed effect (Meltzer 2002), theoretically predictable in all but exceptional circumstances (Linn and Slinde 1977). In view of the threat to equivalence identified earlier, and acknowledging a recognised pitfall of cluster randomisation (Song and Herman 2010), it was necessary to check whether the random process through which schools had been assigned to experimental groups, accompanied by the vicarious process through which schools had chosen participating teachers and classes, had yielded equivalent groups. Unfortunately, this proved not to be the case. For all module topics, the pre-test mean for the experimental group was substantially (and statistically significantly) lower than for the control group. On attitude, while the pre-quest means for mathematics were equivalent, those for science were significantly different.

The accepted response to a confound of this type is to build the predictor variable in as a covariate, although reservations have been expressed that doing so removes variance from the between-condition comparison (Miller and Chapman 2001). To validate such an approach, it was triangulated in the following way. A matching technique was employed to create substantial subsets (around 75%) of the intervention and control samples having almost identical pre-test distributions. Regardless of inclusion or exclusion of pre-test as covariate, the effect sizes produced by these matched-subsample models were almost identical to those produced by full-sample covariate-adjusted models. This triangulation suggests that the results from covariate-adjusted full-sample models provide trustworthy estimates of effect size and significance level.

Accordingly, we report the results of between-condition ANCOVA tests computed using the *Statistical Package for the Social Sciences*. In these tests, the dependent variable was the focal criterion measure of learning gain, attitude change or opinion rating as appropriate, and the independent variables were (i) the corresponding covariate of prior attainment or attitude and (ii) the treatment conditions under comparison – intervention or control group. We report the (Cohen's *d*) effect size of differences between experimental groups as well as their statistical significance.

## Research results

The main results from the field study fall into two parts. First, an observational analysis assesses the degree of implementation of the distinctive pedagogical feature of the intervention, dialogic teaching, focusing on the incidence of markers of classroom dialogic activity. Then a test- and questionnaire-based study evaluates the effectiveness of the *epiSTEMe* intervention in the first year of implementation, focusing on the topic proficiency, module opinion and subject attitude of pupils.

### *Implementation of dialogic teaching*

Assessment of the implementation of dialogic teaching was through analysing the observational evidence from a sample of intervention lessons. Although only one lesson per class was coded, the informal observation that interactions ran smoothly in all of them and that pupils appeared to be familiar with, and proficient in, the kinds of interactions taking place, suggested that these lessons respected established patterns of interaction that participants expected to follow. Table 7 shows the mean incidence of the dialogic markers across all six of the observed lessons for each module, expressed in terms of the proportion of units of observation in which each marker featured. Table 8 shows the lowest incidence observed in any lesson from each module; and where this incidence is 0%, the number of lessons in which this was the case is also given. Because no observations were made of control group classes we cannot know definitively whether these dialogic markers occur less frequently in lessons following established teaching approaches. However, we can examine their incidence within the intervention group.

In terms of overall incidence, teacher solicitation and pupil articulation are prevalent for all topic modules, although levels are rather higher for Probability and Ratios than for Electricity. Attention to multiple pupil ideas is reasonably prevalent in Probability lessons, but notably less so for Electricity and Ratios. Teacher spotlighting of pupil ideas and comparison of ideas are the rarest markers, with appreciable levels only in Probability lessons. In terms of consistency of occurrence, teacher solicitation and pupil articulation occur in almost all lessons on all three topics, but only Probability maintains this consistency for attention to multiple pupil ideas. Teacher spotlighting of pupil ideas and comparison of ideas are absent from virtually all Electricity and Ratios lessons, but present in most Probability lessons. On this basis, the Probability module appears to have been more successful than the others in supporting dialogic teaching, in terms both of generating the full range of dialogic markers and of doing so consistently across lessons.

### *Evaluation of the intervention*

Evaluation of the intervention employed three criterion variables: learning gain (i.e. change in topic proficiency), module opinion and attitude change. In each case, the research question was whether outcomes for pupils in the *epiSTEMe* intervention group differed from those of pupils in the control group.

**Table 7.** Mean incidence of markers of dialogic talk over six lessons, by module: percentage of observational units in which marker occurred.

| | Marker | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Module | TSolC | TSolF | PArtR | PArtE | PMul | TMul | TSpot | PCom | TCom |
| Electricity | 33% | 17% | 36% | 28% | 9% | 9% | 7% | 0% | 0% |
| Probability | 45% | 39% | 48% | 48% | 32% | 27% | 15% | 9% | 13% |
| Ratios | 48% | 46% | 36% | 35% | 11% | 9% | 3% | 0% | 5% |

Table 8. Lowest incidence of markers of dialogic talk in any one lesson, by module: Percentage of observational units in which marker occurred [and number of lessons (out of 6) for which incidence was 0%].

| | Marker | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Module | TSolC | TSolF | PArtR | PArtE | PMul | TMul | TSpot | PCom | TCom |
| Electricity | 11% | 0%[1] | 11% | 0%[1] | 0%[2] | 0%[3] | 0%[4] | 0%[6] | 0%[6] |
| Probability | 30% | 22% | 33% | 33% | 10% | 10% | 0%[2] | 0%[3] | 0%[1] |
| Ratios | 22% | 30% | 11% | 0%[1] | 0%[4] | 0%[4] | 0%[5] | 0%[6] | 0%[5] |

Table 9. Mean learning gain* between pre-test and deferred post-test for each topic, by condition.

| Topic | Intervention gain | Control gain | Relative effect size¶ | Statistical significance† |
|---|---|---|---|---|
| Electricity | +15.6 | +19.3 | −0.20 | $p = .018$ |
| Probability | +10.8 | +9.5 | +0.09 | *ns* |
| Ratios | +7.7 | +4.7 | +0.17 | $p = .033$ |

*In percentage points.
¶Cohen's *d*.
†*F*-test.

### *Pupil learning gain*

We focus on learning gain at deferred post-test as the best indicator of secure progress in topic proficiency. Table 9 presents results for each of the three modules. It is important to note that these figures indicate that mean learning gain was positive whatever the topic or condition. Nevertheless, they also suggest contrasts between modules. Ratios was the only topic for which the intervention group was not advantaged by spending more time on the topic. Here, the effect size +0.17 was statistically significant, a small positive effect. For both the other topics, the intervention group was potentially advantaged by having spent more time on the topic. In the case of Probability, the effect size of +0.09 is not statistically significant. On this basis, the most plausible conclusion is of a null effect. In the case of Electricity, the statistically significant effect size is −0.20. Here the most plausible conclusion is that the effect is a small negative one. In the light of these findings, we can reasonably conclude, that, under the conditions of the field trial, the *epiSTEMe* intervention had differential effects (relative to control) on pupil learning according to module, with effect sizes covering the range from small negative to small positive (although it should be noted that the difference between the two mathematics modules is not itself statistically significant).

### *Pupil attitude change*

As noted earlier, initial attitude towards subject did not differ significantly between experimental groups in mathematics but did so in science. For science, then, a correspondingly adjusted comparison would have been particularly desirable, but was found to be not permissible because of non-homogeneity of regression. Thus, we have had to form a judgement on the basis of unadjusted comparison through ANOVA rather than ANCOVA, as shown in Table 10. In terms of statistical significance, neither of the effect sizes differs from zero, and so we infer a null effect in both subjects. We conclude that, under the conditions of the field trial, the *epiSTEMe* intervention had no effect (relative to control) on pupil attitude towards subject.

### *Pupil module opinion*

The within-condition analyses found that initial subject attitude was also a consistently strong predictor of module opinion, with more positive initial attitude associated with more positive module rating. Table 11 shows the results of correspondingly covariate-adjusted comparisons of pupil module opinion for all three topics. These yield no statistically significant effect for any module. On this basis, we conclude that, under the conditions of the field trial, the *epiSTEMe* intervention had no effect (relative to control) on pupil opinion about their classroom experience.

**Table 10.** Mean attitude change* between start- and end-year for each subject, by condition.

| Topic | Intervention mean | Control mean | Relative effect size¶ | Statistical significance† |
|---|---|---|---|---|
| Science | −0.13 | −0.13 | 0.00 | *ns* |
| Mathematics | −0.16 | −0.08 | −0.09 | *ns* |

*On a scale from −3 [Strong disagreement] to +3 [Strong agreement].
¶Cohen's *d*.
†*F*-test.

**Table 11.** Mean opinion rating* for classroom experience of each topic, by condition.

| Topic | Intervention rating | Control rating | Relative effect size¶ | Statistical significance† |
|---|---|---|---|---|
| Electricity | +0.74 | +0.88 | −0.13 | *ns* |
| Probability | +0.75 | +0.79 | −0.04 | *ns* |
| Ratios | +0.72 | +0.84 | −0.12 | *ns* |

*On a scale from −3 [Strong disagreement] to +3 [Strong agreement].
¶Cohen's *d*.
†*F*-test.

## Discussion and conclusions

### *Limitations of the study and method*

Like most research projects, this one was conducted within constraints of time and resources which limited the capacity of the project team to respond to emergent developments. Equally, like all research projects which involve respectful collaboration with schools and teachers, this one was obliged to accept some compromises to the study protocol and some breakdowns in following it which reflect the complex and demanding circumstances in which school colleagues work. Whatever their origins, however, we are obliged to note here some fundamental limitations of the study. First, the compromise made, in response to requests from schools, to not require participation of teacher pairs in each subject undermined part of the rationale on which the project had designed its professional development. Second, despite such compromises the project was not able to achieve a sample of the size and structure originally anticipated. Third, the plan to assign schools to treatments so as to leave sufficient time for planning and preparation conflicted with the practice in some schools of extremely late decision-making about the timetabling of classes and allocation of teachers. This seems likely to have contributed to the experimental groups proving non-equivalent on the basis of evidence which subsequently became available.[8] Fourth, as in many field studies, despite persistent but sensitive chasing of returns by members of the project team, there was a degree of attrition. Fifth, the rate of attrition was higher within the control group, and while the project team were able to gather some evidence throwing light on the reasons, it has not been possible to exclude the possibility that this biased the achieved sample. Sixth, one of the topic tests proved to be unsatisfactory, leading to the exclusion of this topic from further analysis. Seventh, the decision to use the limited resources available to observe only a sample of intervention lessons means that corresponding information is lacking about the control classes which would have aided interpretation of results. In facing these problems, the project team sought to preserve the integrity of the study by taking appropriate mitigating actions within the resources available. It is on this basis that we consider the results to be worth reporting.

Overall, our experience in carrying out the *episTEMe* project has increased our awareness of the difficulties inherent in conducting large-scale, randomised controlled trials in the real world of school education. In particular, in such research, it is rarely possible to guarantee that recruitment, implementation and assessment procedures will be unaffected by the myriad everyday factors that impinge upon life in schools. Such factors risk compromising measures taken by researchers to randomise and to control in a systematic way. Equally, this study highlights how (unlike perhaps some other fields) the behaviour of control group participants cannot be assumed to lack key features of the behaviours

being promoted in the intervention group. In our view, such complications – and arguably limitations – of the method deserve greater recognition – and deeper attention – in the literature advocating it (e.g. Goldacre 2013).

### *Appraising the epiSTEMe intervention*

Broadly speaking, this evaluation has found that, in its present form, and on the occasion of its first implementation, the *epiSTEMe* intervention produces outcomes no different than would be expected under 'business as usual'. Learning gains in the topic proficiency of pupils are no greater overall than would otherwise have been achieved, although there is some variation between modules. Likewise, the opinions of pupils about their classroom experience and their attitudes towards the two subjects are similar to those associated with established teaching approaches.

An understandable reaction to these findings would be that they are disappointing in view of the investment of time and effort that schools and teachers must make in order to implement the intervention. However, another reaction would be that these results are quite encouraging, indicating that the intervention can be taken on without 'implementation dip' which is often considered a normal accompaniment to such innovation (Fullan 2001). From this perspective, a more optimistic prognosis would be that, as teachers' familiarity and proficiency with innovative features of the intervention grew, it might become more effective. A more pessimistic diagnosis would suggest that lack of an 'implementation dip' might be indicative of the intervention simply being assimilated to established practice through some teachers not implementing dialogic methods as well, or as fully, as intended. A realistic appraisal would probably be that both these processes were in play during the field trial.

In their different ways, these reactions raise the question of what might be done to make the *epiSTEMe* intervention more effective, at first implementation and beyond. Results from the field study suggest areas where the design may need refinement. First, the finding that one topic module (Electricity) produced a small negative effect for learning gain (relative to control) suggests that it requires modification before being used further. Likewise, the finding that two of the topic modules (Electricity and Ratios) failed to support stronger and deeper forms of classroom dialogic activity suggests that they would benefit from revision targeted at strengthening this aspect. Nevertheless, the Probability module which produced the strongest and most consistent implementation of a dialogic teaching approach did not prove particularly effective, at least on this first occasion of implementation. However, some caution is required here: because the study did not ascertain whether the control lessons did indeed differ in dialogic terms, we cannot draw confident conclusions on this point.

Still, conjecturing that, in the longer term, stronger implementation of a dialogic teaching approach would prove more effective raises the question of how best to secure such strengthening. While there is scope to improve the induction provided by the *epiSTEMe* introductory module and the associated professional development, experience over the duration of the project suggests that fostering appropriate forms of in-school support and coaching would be more productive. In particular, during the phases of design and piloting which preceded the field trial, participating teachers gained support through discussing their teaching with colleagues and through feedback from lesson observation by members of the development team.

Not only did project resources not permit such in-school support to be provided during the field trial, but the scale of the professional development associated with the intervention was deliberately limited to reflect conditions currently typical of implementation at scale. We have also noted that a substantial proportion of participating schools chose not to follow the recommendation that pairs of teachers be nominated and so intervention teachers were often working alone. Research suggests that having in-school peer support for implementing change can be an important factor for ensuring success (Dudley 2012; Horn and Kane 2015), and this was clearly not available for many of our participants. However, once some successful practice has developed within a school, it becomes viable to draw on the internal expertise of teachers to support their colleagues. For a school considering implementing

the *epiSTEMe* intervention, the results of the field trial indicate that using the introductory module followed by the relatively successful mathematics modules could be expected, even in the first year, to provide an initial basis for developing staff expertise in dialogic teaching as well as producing a modest overall rise in pupil learning gains.

### *Advancing the research field*

In terms of advancing the research field, this trial of the *epiSTEMe* intervention contributes to the body of evidence about effective teaching in early-secondary science and mathematics through providing results from a systematic, large-scale evaluation of a pedagogical approach incorporating the teaching component of dialogic teaching, a distinctive blend primarily of domain-specific enquiry and cooperative learning. The findings that we have reported suggest that, on first implementation at scale, the *epiSTEMe* intervention is little to no more effective overall in promoting high pupil attainment and favourable pupil attitude than established teaching approaches (although, perhaps as important, no less effective).

In many respects, these findings mirror those of other recent British research on similar lines of pedagogical development. In particular, Osborne et al. (2013) report on a project which sought to develop a more dialogic approach to secondary school science teaching based on small group work and the consideration of ideas, evidence and argument. Evaluation results showed few differences in the conceptual understanding, reasoning and attitudes toward science of pupils in the intervention group compared to the comparison sample. This project described its intervention as based on 'a minimalist programme of professional development and support' (323) offered to two lead teachers in each participating school. Occupying 5 days spread over two school years, the professional development programme made use of research-based video materials focusing on the development of argumentation in the science classroom, and provided opportunities for the lead teachers to share teaching resources and strategies that could be taken back to their science departments and shared with their colleagues. The role of these lead teachers was, in turn, to lead colleagues in their own school department in developing schemes of work and teaching practices which embedded argumentation activities. Reflecting on their project, Osborne et al. suggest that substantive change in teacher knowledge and practices may require more sustained and intensive professional development, and raise the question of whether effects might have been seen in the years after the intervention once teachers had assimilated the practice more fully.

This issue of the limited scale and scope of professional development is equally relevant to the *epiSTEMe* intervention. Relating these British findings back to the existing literature, it is notable that, in the meta-analyses referred to in the earlier literature review, only those reported by Slavin and his colleagues acknowledge the professional development aspect: their overarching construct of 'instructional process programmes' is characterised as covering 'approaches to mathematics reform that emphasise extensive professional development to help teachers use effective teaching strategies' (Slavin, Lake, and Groff 2009, 858). This suggests that the scale and quality of professional development provided to teachers ought to be more explicitly identified as a potentially crucial variable in such meta-analyses. And experience from this project suggests that limited commitment to professional development at both system and school levels may be a major factor inhibiting the successful implementation of potentially more effective teaching practices which call for substantial professional learning.

Finally, the field study has provided some indications of lines of enquiry into dialogic teaching which deserve to be pursued further. One is to further develop observational instruments for dialogic teaching which are both empirically reliable and conceptually cogent. Another is to use such instruments to explore further what may prove to be important relationships between the intensity and depth of dialogic teaching and the effectiveness of learning.

## Notes

1. Extended extracts of classroom talk relating to the illustrative lesson, and analysis of these, are provided in the reference cited (Ruthven, Hofmann, and Mercer 2011).
2. In March, invitations were sent directly to several hundred schools within the area, and a similar invitation was posted prominently on the Faculty website. Two briefing sessions were held in March and April, attended by 22 schools, of which 15 subsequently confirmed participation. In order to increase the sample size, further briefing sessions were held in June and July for smaller numbers of schools which subsequently expressed interest, and 11 further schools were recruited.
3. In early June, all 14 schools which had at that point confirmed their participation were assigned to a condition. For each pair, ordered by CVA score, a computer-generated random number, either 1 or 2, was generated and the school in the corresponding position assigned to the intervention group. By mid-June, a further batch of six schools had confirmed and been assigned. As time pressure increased, further assignments were made as soon as a new pair of confirmations had been received: three further pairs were assigned in this way at points between late June and late July.
4. Because schools made the decision about when a topic would be taught (and indeed whether it would be taught), sometimes doing so at short notice without informing the research team, the scope for targeted reminders was limited. In addition, whereas for the intervention group, using the *epiSTEMe* materials to teach the topic served as a reminder to administer instruments, this was not so for the control group, contributing to the higher rate of attrition in the latter. While it would have been desirable to compare the characteristics of those classes which did not complete the trial with those of the classes that did, unsurprisingly it was often the case that schools failed to provide the basic information requested about non-completing classes and their students which would have made this possible.
5. The deferred post-tests used can be downloaded from <http://www.educ.cam.ac.uk/research/projects/episteme/instruments.html> .
6. A technical report on the observation protocol, including exemplification of each code from the developmental stage, can be downloaded from <http://www.educ.cam.ac.uk/research/projects/episteme/instruments.html> . The preliminary analysis identified seven markers which achieved good levels of inter-observer agreement, and a further two more rarely occurring markers which were judged particularly indicative and achieved acceptable levels of inter-observer agreement. Every observational unit was coded for the presence (or absence) of each marker. Agreement between observers in respect of a marker was assessed in terms of the proportion of units where both observers coded the presence of that marker out of all units where one or both observers coded its presence.
7. Even after chasing by the project team, this questionnaire or simply this particular information was missing from otherwise complete returns in around 20% of cases. Rather than reducing the sample size still further, we retained such cases in the main analysis and conducted analyses of this questionnaire on the basis of the cases for which data were available.
8. In particular, since target topics were taught at the times determined by teachers (normally as scheduled in their school's scheme of work for the subject) which could be late in the school year, a full set of pre-test results was not available until the trial had been completed. Equally, it was not known definitively which classes had failed to return results until that point.

## Acknowledgements

This work was supported by the Economic and Social Research Council which provided funding for the *epiSTEMe* project [grant number RES-179-25-0003]. Thanks are also due to the teachers who generously volunteered to review, pilot and trial versions of the modules, to Christine Howe for her contribution to design and analysis, and to Andy Tolmie and Anna Vignoles for statistical advice.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Notes on contributors

*Kenneth Ruthven* is a professor in the Faculty of Education at the University of Cambridge. His main research interests are in curriculum, pedagogy and assessment, especially in mathematics and particularly in the light of technological change.

*Neil Mercer* is an emeritus professor in the Faculty of Education at the University of Cambridge. His main research interests are in the role of language in the classroom and the development of children's thinking.

*Keith S. Taber* is professor of science education, and current chair of the Science, Technology and Mathematics Education Academic Group, in the Faculty of Education at the University of Cambridge. His research interests mostly cluster around aspects of teaching and learning in science subjects.

*Paula Guardia* was a research associate in the Faculty of Education at the University of Cambridge. She is now an assistant professor of learning and development in the Faculty of Education at the Pontificia Universidad Catolica de Chile. Her particular research interests are in emergent literacy abilities and reading development interventions for children from deprived sociocultural areas.

*Riikka Hofmann* is a research and teaching associate in the Faculty of Education at the University of Cambridge. Her research focuses on interaction in classrooms and professional contexts, learning through talk, pedagogic innovations and professional learning and change in various institutional settings, studied from the perspective of sociocultural psychology.

*Sonia Ilie* is a research fellow in the Faculty of Education at the University of Cambridge. Her research interests include inequalities in access to education, student attainment and progression, educational effectiveness and school leadership and quantitative methods in education.

*Stefanie Luthman* was a research associate in the Faculty of Education at the University of Cambridge. Her research interests include the evaluation of teaching interventions and students' attitudes towards mathematics and science. She is now a senior research analyst at Quantify Research in Stockholm.

*Fran Riga* is a research and teaching associate in the Faculty of Education at the University of Cambridge. Her research interests include conceptual development in science and associated thinking processes, education for the gifted in science (especially females), dialogic approaches to teaching and learning, inquiry-based science education and adaptive learning.

## ORCID

*Neil Mercer* http://orcid.org/0000-0002-6829-8072

## References

Bennett, J., S. Hogarth, F. Lubben, B. Campbell, and A. Robinson. 2010. "Talking Science: The Research Evidence on the Use of Small Group Discussions in Science Teaching." *International Journal of Science Education* 32 (1): 69–95.

Bennett, J., F. Lubben, S. Hogarth, and B. Campbell. 2005. "Systematic Reviews of Research in Science Education: Rigour or Rigidity?" *International Journal of Science Education* 27 (4): 387–406.

Black, P., and J. M. Atkin. 2014. "The Central Role of Assessment in Pedagogy." In *Handbook of Research on Science Education*. Vol. II, edited by N. G. Lederman and S. K. Abell, 775–790. New York: Routledge.

Bransford, J., A. Brown, and R. Cocking. 2000. *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academies Press.

Davis, E., and J. Krajcik. 2005. "Designing Educative Curriculum Materials to Promote Teacher Learning." *Educational Researcher* 34 (3): 3–14.

Department for Education. 2010. *Key Stage 2 to Key Stage 4 (KS2-KS4): Contextual Value Added Measure (CVA) including English and Maths*. Accessed February 20, 2015. http://www.education.gov.uk/schools/performance/archive/schools_10/s3.shtml

DfEE (Department for Education and Employment). 1998. *The Implementation of the National Numeracy Strategy: The Final Report of the Numeracy Task Force*. London: DfEE.

DfEE (Department for Education and Employment). 2001. *Key Stage 3 National Strategy: Framework for Teaching Mathematics: Years 7, 8 and 9*. London: DfEE.

DoE (Department of Education). 1999. *Exemplary and Promising Mathematics Programs*. Washington, DC: DoE.

Dimitrov, D., and P. Rumrill. 2003. "Pretest–Posttest Designs and Measurement of Change." *Work: A Journal of Prevention, Assessment and Rehabilitation* 20 (2): 159–165.

Dudley, P. 2012. "Lesson Study Development in England: From School Networks to National Policy." *International Journal for Lesson and Learning Studies* 1 (1): 85–100.

Duschl, R., H. Schweingruber, and A. Shouse. 2007. *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, DC: National Academies Press.

Fullan, M. 2001. *Leading in a Culture of Change*. San Francisco, CA: Jossey-Bass.

Goldacre, B. 2013. *Building Evidence into Education*. London: Department for Education.

Good, T. L., D. A. Grouws, and H. Ebmeier. 1983. *Active Mathematics Teaching*. New York: Longman.

Horn, I. S., and B. D. Kane. 2015. "Opportunities for Professional Learning in Mathematics Teacher Workgroup Conversations: Relationships to Instructional Expertise." *Journal of the Learning Sciences* 24 (3): 373–418.

Howe, C., S. Ilie, P. Guardia, R. Hofmann, N. Mercer, and F. Riga. 2015a. "Principled Improvement in Science: Forces and Proportional Relations in Early Secondary-school Teaching." *International Journal of Science Education* 37 (1): 162–184.

Howe, C., S. Luthman, K. Ruthven, N. Mercer, R. Hofmann, S. Ilie, and P. Guardia. 2015b. "Rational Number and Proportional Reasoning in Early Secondary School: Towards Principled Improvement in Mathematics." *Research in Mathematics Education* 17 (1): 38–56.

Howe, C., and A. Tolmie. 2003. "Group Work in Primary School Science: Discussion, Consensus and Guidance from Experts." *International Journal of Educational Research* 39 (1–2): 51–72.

Howe, C., A. Tolmie, A. Thurston, K. Topping, D. Christie, K. Livingston, E. Jessiman, and C. Donaldson. 2007. "Group Work in Elementary Science: Towards Organisational Principles for Supporting Pupil Learning." *Learning and Instruction* 17 (5): 549–563.

Kilpatrick, J. 2012. "The New Math as an International Phenomenon." *ZDM* 44 (4): 563–571.

Kilpatrick, J., J. Swafford, and B. Findell. 2001. *Adding it up: Helping Children Learn Mathematics*. Washington, DC: National Academy Press.

Kyriacou, C., and J. Issitt. 2008. *What Characterises Effective Teacher-initiated Teacher-pupil Dialogue to Promote Conceptual Understanding in Mathematics Lessons in England in Key Stages 2 and 3: A Systematic Review*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Lemke, J. L. 1990. *Talking Science: Language, Learning, and Values*. Norwood, NJ: Ablex Publishing Corporation.

Linn, R. L., and J. A. Slinde. 1977. "The Determination of the Significance of Change between Pre- and Posttesting Periods." *Review of Educational Research* 47 (1): 121–150.

Louis, K. S., H. M. Marks, and S. Kruse. 1996. "Teachers' Professional Community in Restructuring Schools." *American Educational Research Journal* 33 (4): 757–798.

Meltzer, D. E. 2002. "The Relationship between Mathematics Preparation and Conceptual Learning Gains in Physics: A Possible "Hidden Variable" in Diagnostic Pretest Scores." *American Journal of Physics* 70 (12): 1259–1268.

Mercer, N., L. Dawes, R. Wegerif, and C. Sams. 2004. "Reasoning as a Scientist: Ways of Helping Children to Use Language to Learn Science." *British Educational Research Journal* 30 (3): 367–385.

Mercer, N., and K. Littleton. 2007. *Dialogue and the Development of Children's Thinking: A Sociocultural Approach*. London: Routledge.

Mercer, N., and C. Sams. 2006. "Teaching Children How to Use Language to Solve Maths Problems." *Language & Education* 20 (6): 507–527.

Miller, G. A., and J. P. Chapman. 2001. "Misunderstanding Analysis of Covariance." *Journal of Abnormal Psychology* 110 (1): 40–48.

National Academy of Sciences. 1995. *National Science Education Standards*. Washington, DC: National Academies Press.

NCTM (National Council of Teachers of Mathematics). 2000. *Principles and Standards for School Mathematics*. Reston, VA: NCTM.

OfStEd (Office for Standards in Education). 2008a. *Success in Science*. London: OfStEd.

OfStEd (Office for Standards in Education). 2008b. *Mathematics: Understanding the Score*. London: OfStEd.

Ogborn, J., G. Kress, I. Martins, and K. McGillicuddy. 1996. *Explaining Science in the Classroom*. Buckingham: Open University Press.

Osborne, J., S. Simon, A. Christodoulou, C. Howell-Richardson, and K. Richardson. 2013. "Learning to Argue: A Study of Four Schools and Their Attempt to Develop the Use of Argumentation as a Common Instructional Practice and Its Impact on Students." *Journal of Research in Science Teaching* 50 (3): 315–347.

Osborne, J., S. Simon, and S. Collins. 2003. "Attitudes towards Science: A Review of the Literature and Its Implications." *International Journal of Science Education* 25 (9): 1049–1079.

Plewis, I., and J. Hurry. 1998. "A Multilevel Perspective on the Design and Analysis of Intervention Studies." *Educational Research and Evaluation* 4 (1): 13–26.

Reynolds, D., and R. D. Muijs. 1999. "The Effective Teaching of Mathematics: A Review of Research." *School Leadership and Management* 19 (3): 273–288.

Ruthven, K. 2011. "Using International Study Series and Meta-analytic Research Syntheses to Scope Pedagogical Development Aimed at Improving Student Attitude and Achievement in School Mathematics and Science." *International Journal of Science and Mathematics Education* 9 (2): 419–458.

Ruthven, K., and R. Hofmann. 2013. "Chance by Design: Devising an Introductory Probability Module for Implementation at Scale in English Early-secondary Education." *ZDM* 45 (3): 409–423.

Ruthven, K., R. Hofmann, C. Howe, S. Luthman, N. Mercer, and K. S. Taber. 2012. "The *EpiSTEMe* Pedagogical Approach: Essentials, Rationales and Challenges." *Proceedings of the British Society for Research into Learning Mathematics* 31 (3): 131–136.

Ruthven, K., R. Hofmann, and N. Mercer. 2011. "A Dialogic Approach to Plenary Problem Synthesis." *Proceedings of the 35th Conference of the International Group for the Psychology of Mathematics Education* 4: 81–88.

Scheerens, J., M. Hendriks, H. Luyten, P. Sleegers, and C. Glas. 2013. *Productive Time in Education. A Review of the Effectiveness of Teaching Time at School, Homework and Extended Time outside School*. Accessed February 20, 2015. http://doc.utwente.nl/86371/1/Productive_time_in_education.pdf

Schroeder, C. M., T. P. Scott, H. Tolson, T.-Y. Huang, and Y.-H. Lee. 2007. "A Meta-analysis of National Research: Effects of Teaching Strategies on Student Achievement in Science in the United States." *Journal of Research in Science Teaching* 44 (10): 1436–1460.

Scott, P. H. 1998. "Teacher Talk and Meaning Making in Science Classrooms: A Vygotskian Analysis and Review." *Studies in Science Education* 32: 45–80.

Scott, P. H., H. M. Asoko, and J. Leach. 2007. "Student Conceptions and Conceptual Change Learning in Science." In *Handbook of Research on Science Education*, edited by S. Abell and N. Ledermann, 31–56. Mahwah, NJ: Erlbaum.

Scott, P. H., E. F. Mortimer, and O. G. Aguiar. 2006. "The Tension between Authoritative and Dialogic Discourse: A Fundamental Characteristic of Meaning Making Interactions in High School Science Lessons." *Science Education* 90 (4): 605–631.

Seidel, T., and R. J. Shavelson. 2007. "Teaching Effectiveness Research in the past Decade: The Role of Theory and Research Design in Disentangling Meta-analysis Research." *Review of Educational Research* 77 (4): 454–499.

Slavin, R., and C. Lake. 2008. "Effective Programs in Elementary Mathematics: A Best-evidence Synthesis." *Review of Educational Research* 78 (3): 427–515.

Slavin, R., C. Lake, and C. Groff. 2009. "Effective Programs in Middle and High School Mathematics: A Best-evidence Synthesis." *Review of Educational Research* 79 (2): 839–911.

Song, M., and R. Herman. 2010. "Critical Issues and Common Pitfalls in Designing and Conducting Impact Studies in Education: Lessons Learned from the What Works Clearinghouse (Phase I)." *Educational Evaluation and Policy Analysis* 32 (3): 351–371.

Supovitz, J. A., and H. M. Turner. 2000. "The Effects of Professional Development on Science Teaching Practices and Classroom Culture." *Journal of Research in Science Teaching* 37 (9): 963–980.

Taber, K. S., K. Ruthven, C. Howe, N. Mercer, F. Riga, R. Hofmann, and S. Luthman. 2015. "Developing a Research-informed Teaching Module for Learning about Electrical Circuits at Lower Secondary School Level: Supporting Personal Learning about Science and the Nature of Science." In *Cases on Research-based Teaching Methods in Science Education*, edited by E. de Silva, 122–156. Hershey, PA: IGI Global.

Trafton, P., B. Reys, and D. Wasman. 2001. "Standards-based Mathematics Curriculum Materials: A Phrase in Search of a Definition." *Phi Delta Kappan* 83 (3): 259–264.