*Research Article*

# 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

Jonathan Osborne,[1] Shirley Simon,[2] Andri Christodoulou,[3] Christina Howell-Richardson,[4] and Katherine Richardson[2]

[1]*Graduate School of Education, Stanford University, 485 Lasuen Mall, Stanford, California 94305*
[2]*Institute of Education, University of London, London, UK*
[3]*University of Southampton, Southampton, UK*
[4]*King's College London, School of Education, London, UK*

*Received 23 January 2012; Accepted 6 December 2012*

Abstract: This article reports the outcomes of a project in which teachers' sought to develop their ability to use instructional practices associated with argumentation in the teaching of science—in particular, the use of more dialogic approach based on small group work and the consideration of ideas, evidence, and argument. The project worked with four secondary school science departments over 2 years with the aim of developing a more dialogic approach to the teaching of science as a common instructional practice within the school. To achieve this goal, two lead teachers in each school worked to improve the use of argumentation as an instructional practice by embedding activities in the school science curriculum and to develop their colleague's expertise across the curriculum for 11- to 16-year-old students. This research sought to identify: (a) whether such an approach using minimal support and professional development could lead to measurable difference in student outcomes, and (b) what changes in teachers' practice were achieved (reported elsewhere). To assess the effects on student learning and engagement, data were collected of students' conceptual understanding, reasoning, and attitudes toward science from both the experimental schools and a comparison sample using a set of standard instruments. Results show that few significant changes were found in students compared to the comparison sample. In this article, we report the findings and discuss what we argue are salient implications for teacher professional development and teacher learning. © 2013 Wiley Periodicals, Inc. J Res Sci Teach 50:315–315, 2013

**Keywords:** teacher professional development; argumentation; dialogic teaching; student engagement; epistemic learning

Within research in science education over the past decade, the study of argumentation has been a prominent feature. The case for its role and significance in the learning of science was made initially by Kuhn (1993) and Driver, Newton, and Osborne (2000). Since then work in this sphere of research has expanded substantially. For instance a search of the terms "argumentation," "science," and "education" in the ERIC database shows that, while there were only 34 publications between 2000 and 2004, 84 have been published between 2005

Additional Supporting Information may be found in the online version of this article.
*Correspondence to:* Jonathan Osborne; E-mail: osbornej@stanford.edu
DOI 10.1002/tea.21073
Published online 17 January 2013 in Wiley Online Library (wileyonlinelibrary.com).

and 2009. Likewise, in an analysis of research trends in science education from 2003 to 2007, Lee, Wu, and Tsai (2009) found that 4 of the 10 most highly cited articles from 1998 to 2002 included the word "argumentation" in their title, and between 2003 and 2007, 5 of the 10 most highly cited articles included either the word "argumentation" or "informal reasoning" in their title. In addition, within science education, two edited volumes exist on argumentation (Erduran & Jiménex-Aleixandre, 2008; Khine, 2011). The interest is not confined to science alone with work on argumentation in mathematics (e.g., Ball & Bass, 2000; Boaler & Staples, 2008; Cobb et al., 1991; Lampert, 1990; Stein, Engle, Smith, & Hughes, 2008), language arts (Andrews, 1995; Applebee, Langer, Nystrand, & Gamoran, 2003; Nystrand, Gamoran, Kachur, & Prendegarst, 1997), and history (Goldberg, Schwarz, & Porat, 2008; Pontecorvo, 1987; Pontecorvo & Girardet, 1993).

From an educational perspective, much of the work has been driven by a socio-cultural perspective (Wertsch, 1991) which sees knowledge as an entity which is co-constructed by "communities of learners" (Brown & Campione, 1994; Greeno, 1998; Sfard, 2008). The central feature of any pedagogic enactment of this theoretical approach is the opportunity for students to engage in dialogic exploration of ideas and their justification—activities which facilitate a dialectic between construction and critique (Ford, 2008). Such a vision is possibly articulated most strongly by Alexander (2005) who argues that:

> Children construct meaning not only from the interplay of what they newly encounter and what they already know, but also from the interaction with others. In turn, this interaction is critical not just for children's understanding of the kind of knowledge with which schools deal—mathematics, science and so on—but also for the development of their very identity, their sense of self and worth (p. 11).

Alexander defines his conception of the dialogic classroom as one which is *collective*—in that teachers and children address learning tasks together; *reciprocal* in that teachers and children listen to each other and consider alternative viewpoints; *supportive* in that children articulate their ideas freely helping each other to reach common understandings; *cumulative* in that teachers and children build on their own and each others' ideas; and *purposeful* in that teachers plan and facilitate dialogic teaching with well-defined educational goals in view. Similar ideas are the foundation of the concept of "accountable talk" where talk is required to be "accountable to the community," "accountable to standards of reasoning," and "accountable to knowledge" (Michaels, O'Connor, & Resnick, 2008). Such characterizations see engaging in deliberative discourse and argumentation as something that is central to learning.

Indeed, emerging from the work of social psychologists is increasing empirical evidence that the knowledge and understanding of school-age children can be facilitated by collaborative dialog between peers (Alverman, Qian, & Hynd, 1995; Damon & Phelps, 1989; Doise & Mugny, 1984; Howe, Tolmie, & Mackenzie, 1995; Mercer, Dawes, Wegerif, & Sams, 2004; Venville & Dawson, 2010; Zohar & Nemet, 2002). For instance, the work of Howe et al. (1995), and Alverman et al. (1995) with young students has demonstrated that a consideration of plural perspectives about a scientific phenomenon leads to an improved and more secure conceptual understanding—that is knowing why the wrong idea is wrong matters as much as knowing why the right idea is right. Perhaps the strongest evidence for the value of a dialogic approach to learning comes from a review conducted by Chi of empirical evidence from studies of children's learning (Chi, 2009). She developed a taxonomy of activities to be undertaken by learners as either *interactive*—defined as any activity where learners interact with peers in joint dialogue with the goal of constructing a product; or *constructive* in that they require learners to produce individually some form of overt output that is not contained within the materials to hand; or *active* in the sense that students are physically doing something whilst learning. Students who are simply watching a video, for instance, would be considered to be at most active in that, at best, it requires the students' cognitive engagement. Chi then searched the literature for empirical studies that made any relevant pairwise comparisons, for example, interactive versus active, constructive versus active, interactive versus constructive to establish that interactive approaches were significantly more effective for learning than constructive approaches which, in turn, were significantly more effective than active approaches.

The theoretical position adopted in this research is that developing student understanding requires the juxtaposition of competing alternatives—often between the students' pre-existing conception and the newly presented scientific idea. Argumentation is then a vital means of generating cognitive conflict forcing the student to identify their current conception and engage in the cognitive act of comparison, contrast, and evaluation of evidence. Confronted with the uncertainty caused by the need to resolve two conflicting ideas, students like all humans resort to reasoning which is fundamentally probabilistic. Innovative descriptions of this process have been developed by using Bayesian models of human reasoning (Howson & Urbach, 2006; Oaksford & Chater, 2007) where knowledge construction is perceived as a dialectic requiring the evaluation of competing alternatives (Szu & Osborne, 2011). As such, the process of knowledge construction can be seen as being dependent on *learning to argue* (Billig, 1996). The primary distinction between scientific reasoning and other forms of reasoning is the domain-specific forms of knowledge, that is, conceptual, procedural, and epistemic, on which any argument must draw.

## Background and Rationale

### *Enhancing Scientific Reasoning and Conceptual Understanding*

As explained above, for the purpose of this research, reasoning was operationalized as the ability to construct and critique arguments about scientific phenomena. For construction this can be seen as the ability to identify claims, warrants and data, and the relationship between these elements. Included as well was the ability to coordinate theory and evidence—that is to construct warrants that relate data to an appropriate hypothesis. Critique likewise requires the ability to identify the elements of an argument but, in addition, the ability to evaluate the validity the data and warrants of competing alternatives.

In working with the teachers and schools that were the focus of this study, therefore, we sought to develop instructional practices that would promote argumentative and deliberative discourse building on what we had learnt from earlier projects in this domain (Osborne, Erduran, & Simon, 2004a, 2004b; Simon, Erduran, & Osborne, 2006). A more dialogic approach to pedagogy requires the ability to facilitate discussion among students, identify the features of an argument, model argument for students, encourage students to take a position, defend it with evidence-based arguments, and evaluate and generate appropriate counter-arguments. The kind of discourse and reasoning we sought to promote is similar to that to be found in Michaels et al. (2008) notion of "accountable talk" which uses a set of prototypical discursive moves such as asking students to restate someone's reasoning; asking students' to apply their own reasoning to someone else's reasoning; prompting students for further participation; asking students to explicate their reasoning; and offering challenges and counter-examples. The emphasis of such discourse is on the elaboration of reasoning rather than simply supporting or criticizing conclusions. While scientific reasoning is dependent on scientific knowledge as Resnick, Michaels, and O'Connor (in press) point out:

> Good reasoning, hence good discourse, depends on good knowledge. At the same time, acquiring good knowledge depends on active processing and good reasoning. Knowledge and reasoning develop best in tandem; neither precedes the other.

To date, there is an accumulating body of studies which suggest that engaging in argumentation can lead to significant improvements in students' conceptual understanding when compared to students taught by other methods (Mercer et al., 2004; Venville & Dawson, 2010; Zohar & Nemet, 2002). However, the notable feature of all of these studies is that they are essentially what might be termed "proof of concept." That is small-scale studies using volunteer teachers to enact instructional practices that use argumentation with the goal of improving student learning of scientific concepts. What is missing from this work has been any attempt to explore whether it is possible to embed the use of argumentation into the daily practices of teaching science with larger, and perhaps less-willing groups of science teachers—that is to make it a "form of life" which students and teachers engage in on a regular basis. What kind of professional development, we asked would it require to change the cultural habitus of a group of teachers working together in a school such that opportunities to engage in deliberative discourse of ideas and evidence would be a more prominent feature of the classroom? And, moreover, what effect would that have on student engagement with learning and its outcomes? It is the latter question which is the focus of the research reported here.

### *Model of Professional Development*

Using a dialogic approach is a practice which is neither familiar nor comfortable for many science teachers and can pose a significant challenge to their teaching (Simon et al., 2006; Zohar, 2004). Transforming pedagogy requires teachers to share the values of an innovation and be prepared to take risks (Claxton, 1988; Loucks-Horsley, Hewson, Love, & Stiles, 2003)—a venture that is best supported by establishing the practice of collaborative reflection within a supportive community focused on professional learning (Hoban, 2002). Early approaches to teacher development that had little sustained impact were underpinned by mistaken beliefs that teacher learning is a linear process where teachers' practice could be transformed by prescriptive approaches. In contrast, current knowledge would suggest that a more complex view of professional learning is required to bring about sustained change (Desimone, 2009; Opfer & Pedder, 2011).[1] Hoban's work is particularly important in identifying a combination of conditions for teacher learning that complement each other in supporting change. These are a conception of teaching as a dynamic relationship with students and with other teachers where change involves uncertainty; room for reflection in order to understand the emerging patterns of change; a sense of purpose that fosters the desire to change; a community to share experiences; opportunities for action to test what works or does not work in their classrooms; conceptual inputs to extend teachers' knowledge and experience (in this case, ideas about the value of argumentation in teaching science); and finally sufficient time to adjust to the changes made. Similar findings emerge from a comprehensive analysis of 1,300 studies on effective professional development (Yoon, Ducan, Lee, Scarloss, & Shapley, 2007) and from the work of Desimone (2009).

However, evidence of change is sparse—particularly when it comes to answering the question of what effect teacher professional development might have on student learning. Yoon and his colleagues could only find nine studies that met the standards of credible evidence specified by the What Works Clearinghouse and all these studies were of professional development in elementary schools. These nine efforts relied on a mix of workshops, inputs from external experts, a minimum of 30 contact hours, structured and sustained follow-up, opportunities to adapt varied instructional practices to their own context, and a specific focus on content or pedagogic practices. These nine studies showed that the students of these teachers improved their performance by 21 percentile points when compared to a sample of comparison students. Of these studies, only those that had more than 14 hours of professional development showed a positive effect on student achievement. Likewise, Desimone's review of the field argues that the research literature points to five similar factors which are "critical to increasing teacher knowledge and skills and improving their practice, and which hold promise for increasing student achievement: (a) content focus, (b) active learning, (c) coherence, (d) duration, and (e) collective participation" (p. 183).

Building on this literature, and working with whole school science departments, we sought to investigate whether a cycle of collaborative reflective professional development involving distributed leadership (Spillane, 2006) would enable science teachers in a department to develop a pedagogic practice that made greater use of argumentation in their teaching. Our approach to this challenge was essentially both systemic and minimalist. It was systemic in that we sought to work with whole school faculty through two teachers who were to act as instructional leaders. The work would have a focus on developing the use of a specific pedagogic practice—that is the use of argumentation as a means of building student understanding of science content. The work would be active in that it would involve teachers in discussing and reflecting on their practice collectively; it would be coherent in that its focus had a well-defined focus—the use of argumentation in teaching science; and it would be of extended duration in that it would be a collective effort within the department involving teachers in a minimum of 10 monthly meetings of 1–2 hours duration.

The work was minimalist, however, in that we sought to see if change could be achieved by a limited investment in resources to support this process. The rational for the latter choice is essentially economic. Substantial resources are spent by all countries on professional development. For instance, Birman et al. (2007) reckon that the US spent a sum of $1.5 billion at local, state and federal levels on professional development in the period 2004–2005. Investing in professional development, given the thin evidence base for its effectiveness could be seen as a triumph of hope over reality. Our interest lay therefore in asking whether any change in teacher practice and in student learning, could be achieved with a more minimal form of approach to teacher professional development? The model used was a distributed model based on providing 30 hours of professional development to the teacher leaders, meeting the criteria about what is known for minimum duration. These, in turn, were expected to provide approximately 10–14 hours of professional development to each of the teachers in their school. While this would be less than that suggested by research, the question we asked was whether our proposed structure where the whole community of science teachers in schools worked on one specific practice would lead to notable change in their practice, and then, in their students learning and engagement?

In addition, our approach to teacher learning and teacher change involved teachers who are perhaps less motivated to change or try something new. As Desimone (2009) points out, few studies to date have worked with teachers who are not volunteers. Thus, by working with a department, we sought to work with a cohort of teachers who might be more representative of the teachers likely be asked to enact innovative practices. Would it, we asked, be possible to embed the use of an instructional practice within the departmental habitus where it would be seen as a common and valued practice such that any new teachers would be encultured into this instructional practice as the common "form of life" within that institution? Testing such a model is important as policymakers, school superintendents, and school principals work within tight economic constraints often with teachers who see little need to adapt or develop their practice (Dillon, Osborne, Fairbrother, & Kurina, 2000; McLaughlin & Talbert, 1993; Spillane, 1999).

Our approach did, nevertheless, require us to recruit schools as volunteers. To that extent the heads of science departments were volunteers, and recognizing the importance of school principals to the success or failure of the design (Borko, 2004), both PIs visited the principals to explain the importance of the project, its goals, the commitment and the possible outcomes. As a consequence, support for the work was sought at the highest level within the schools. After an initial introduction about the project for faculty, the value of participating was determined by subsequent internal discussions amongst the teachers of science.

Recognizing that any such change might take considerable time to develop the work with the schools was conducted over a period of 2 years which was regarded as the minimum necessary to develop a substantive change towards a more dialogic practice. Using argumentation as a common pedagogic strategy, therefore, must be seen as the acquisition of a new practice. Our theory of action draws on the research on the acquisition of expert performance which points to the need for explicit instruction beginning with a set of simple tasks which then transition to become more complex and challenging. Progress is reliant on formative feedback and the use of practice activities that the individual can use between meetings (Ericsson, Krampe, & Tesch-Römer, 1993). In an ideal world, this is best achieved through individual supervision. In our context, we sought to establish such support through a requirement that schools engage in regular meetings that required *collective reflection on practice* and the sharing of their experiences when using argumentation. The process of reflection itself has become an integral part of how we understand teacher learning and the provision of effective professional development. It occurs in many models of teacher development, often as a fundamental process for stimulating change. In particular, it is a feature of Clarke and Hollingsworth's Interconnected Model (2002) that formed the basis of our intervention. In this model change is seen to occur through the mediating processes of reflection and enactment - a good example of which can be found Sisk-Hilton's work and her use of a "consultancy protocol" where teachers share experiences and knowledge (Sisk-Hilton, 2009). Initiating the kind of change that was attempted in this project was therefore reliant on teachers trying out new approaches, sharing their experiences and reflecting on their own practice. Change was seen to be a complex cyclical interaction between values, beliefs, and experiences of using new strategies in the classroom through repeated deliberate practice. However, new practices threaten the feelings of competence, confidence, control, and comfort that most individuals prefer (Claxton, 1988) and, therefore are not automatically enjoyable. As a consequence, individuals must perceive some long-term benefit that skilled use of the practice might afford. To this end, presentations made to teachers emphasized the potential benefit of this practice to student achievement, engagement, and ability to reason drawing on existing research evidence.

The fact that the schools had agreed to participate reflected a shared belief in the importance of the work by the principals and their head of departments. The same could be said for the two lead teachers in each of the schools and data on their views was collected through three interviews with each of these teachers at the beginning, middle, and end of the project. Less clear, however, is the extent to which the individual teachers within the department shared the values or goals of the project. It is such a context, however, in which any attempt to develop teacher practice must be tested if the results are to have consequential validity. Thus, by using a minimalist intervention of asking all science teachers in one school to commit to the use of an innovational practice, and by supporting their practice by offering collegial support from two informed and trained colleagues who took the role of instructional leaders, we asked whether it would be possible to transform the institutional habitus to one which saw a greater role and value for a more dialogic approach to pedagogy and what effect that might have on the cognitive and affective outcomes for their students?

The latter question is important as changing teacher practice in and of itself is of little value as the goal of professional development is ultimately to improve student learning, motivation, and engagement. Thus the question of whether teacher professional development or teacher learning has any effect on student learning is a critical question for the field (Desimone, 2009; Yoon et al., 2007). In the contemporary context, many interpret this as whether the professional development has a positive effect solely on the cognitive outcomes of student learning. In our work, however, for reasons we now discuss, we sought to examine whether the intervention had an effect not only on students' conceptual understanding but also, their epistemic knowledge and their attitudes towards and engagement with school science.

### *Enhancing Students' Reasoning and Conceptual Knowledge*

As previously mentioned, considerable evidence exists that engaging in argumentation can lead to an enhanced conceptual understanding. Our view is that this outcome is achievable because scientific ideas are evaluated not in isolation but by comparison to alternatives, for example, day and night is caused by a spinning Earth versus day and night is caused by a moving Sun. From a Bayesian perspective, this requires the evaluation of the likelihood of one hypothesis versus the other. This view is consistent with the scientific process as described by Kuhn in his seminal work *The Structure of Scientific Revolutions* (1970): "The decision to reject one paradigm is always simultaneously the decision to accept another, and the judgment leading to that decision involves the comparison of both paradigms with nature and each other" (p. 77). Other authors have made similar observations, such as Graff and Birkenstein (2007) in their characterization of how academic arguments are formulated: "Effective persuasive writers do more than make well-supported claims [. . .]; they also map those claims relative to the claims of others" (p. xix). Indeed scientific manuscripts often focus as much on lowering the likelihood of alternative theories as on raising the likelihood of the new proposed theory—a good example of which is to be found in the Watson and Crick (1953) article for the structure of DNA where the authors begin, not by making the case of their new hypothesis, but by critiquing the existing alternative models. Likewise, Hynd and Alvermann (1986) found that physics texts that containing *refutation text* addressing common misconceptions resulted in significantly better conceptual gains. Similarly, Ames and Murray (1982) found greater learning gains among discussion groups with differing preconceptions versus those that were more congruent, even if those differences were based on incorrect premises. Thus, our hypothesis was that providing space to evaluate competing ideas and to engage in argument from evidence would both require students to reason critically—a premise which is addressed by question 1a—and should lead to an enhanced and more secure understanding of scientific concepts—a hypothesis tested by research question 1b.

### *Enhancing Students Epistemic Knowledge*

Work to date, would suggest that the focus of school science on well-established knowledge leaves students with simplistic positivist notions that scientific knowledge is unequivocal and absolute (Driver, Leach, Millar, & Scott, 1996). Thus, another argument for the value of an instructional approach that placed more emphasis on reasoning and argumentation is that it would enhance students' knowledge and understanding of the epistemic basis of science.

For instance, Driver et al. (2000) argued in their article that engaging in argumentation would lead to a "clarification of the norms by which scientists make rational decisions between alternative hypotheses and begin the process of *establishing the norms of scientific argumentation*[2] (p. 300)."

Research on students' understanding of the epistemology of science itself (as opposed to their personal epistemology) has suggested that students' knowledge and understanding is fairly limited (Driver et al., 1996; Lederman & O'Malley, 1990; Sandoval, 2003). Driver et al. (1996) concluded that students have difficulty determining the role of theories in science and the way that theories are evaluated against existing data by experimentation. In particular, scientific theories were viewed as simple descriptions of taken-for-granted facts, especially for 9-year olds. Moreover for these students, theories were seen to be of lower epistemic status than "facts" which were considered to be the ultimate goal of science. Such a finding stands in stark contrast to the view *within* science that "theories are the crown of science, for in them our understanding of the world is expressed" (p. 168) (Harré, 1984). Views of scientific theories as ideas based in evidence, or as a way of modeling phenomena and predicting their behavior, were found more often in older students' responses. Sandoval and Morrison (2003) have also explored high school students' ideas of scientific theories and theory change finding: (a) that scientific theories were viewed as hypotheses which are repeatedly proven; and (b) that a correspondence view of science predominates—the latter being the conception that scientific theories provide explanations of the world as it is rather than being a map or representation of reality which has the potential to be improved. An important task for science education, therefore, is to develop children's ability to understand that argument and uncertainty are elements that are inherent to science, and that argument itself is a normative practice. To date, a search of the literature would suggest that only five studies have explored the effect of argumentation on middle or high school students' epistemic understanding, two of which have been undertaken by Sandoval and his collaborators (Sandoval & Millwood, 2005; Sandoval & Morrison, 2003). These studies suggest that an enhanced understanding of scientific epistemology supports student engagement in argumentation but only Sandoval and Morrison have looked at the converse—whether engaging in argumentation improves students' epistemic knowledge and their work produced equivocal findings (Sandoval & Morrison, 2003).

More recently, *Taking Science to School* (National Research Council, 2007) and the framework for the common core standards for science in the US (National Research Council, 2012) suggest that engaging in a limited set of scientific practices such as modeling, analyzing, and interpreting data, and constructing explanations are an opportunity to build student understanding of how these practices contribute to the construction of knowledge in science. However, there is a lack of empirical data which might support such arguments leading us to ask whether engaging in practices which require scientific reasoning and argumentation does lead to an enhanced understanding of the epistemology of science? It is this question that forms research question 1c in our study.

### *Engaging Students With Science*

One of the enduring concerns in science education is that the standard form of practice leads to negative attitudes towards science for too large a proportion of students (Cerini, Murray, & Reiss, 2003; Osborne & Collins, 2001). For instance, an analysis of the PISA items on interest in science embedded in the PISA 2006 test shows a clear negative correlation between interest and attainment (Avvisati & Vincent-Lancrin, in press). Such findings suggest that the passion for science may be extinguished by discourse which bears the stamp of authority. As Horton illuminated in his scholarly analysis of the difference between African traditional forms of thought and the rationality of Western science, the school science classroom attends to what we know at the expense of how we know (Horton, 1967). Several recent studies of students' attitude to science would suggest that the student experience of school science is dominated by the perception that school science is an authoritative body of knowledge with a pedagogy dominated by copying and rote learning. For instance, in a study by Weiss, Pasley, Sean Smith, Banilower, and Heck (2003), two-thirds of lessons were graded low quality as they either lacked opportunities for sense making, failed to use teacher questioning to enhance understanding, or lacked intellectual rigor. Not surprisingly, such instructional approaches do little to promote student engagement providing few opportunities for the vital ingredients of autonomy and control (Csikszentmihalyi & Schneider, 2000; Paris, 1998; van Lier, 1996). As a consequence, science is perceived as a discipline that offers little opportunity for creativity or personal realization (Hannover & Kessels, 2004; Taconis & Kessels, 2009). The question then arises whether the use of a more dialogic approach might stimulate student interest by showing that constructing an understanding of the material world is both a creative and challenging intellectual endeavor, as engaging in argumentation requires the presentation of plural alternatives for consideration challenging the notion that there is a singular correct answer. Moreover, it requires the coordination between theory and evidence foregrounding the idea that scientific knowledge is neither self-evident nor easily acquired. Students might then sense that science does offer opportunities for autonomy and ingenuity, which research shows are key factors in generating intrinsic as opposed to extrinsic motivation (Dweck, 2000; Paris, 1998; Renninger, Hidi, & Krapp, 1992). In addition, these two elements are key to developing engagement and a sense of "flow" (Csikszentmihalyi & Hermanson, 1995).

As a consequence, the question we asked was whether the use of a more dialogic approach in the teaching of science would lead to enhanced student engagement with science (research question 1d)? This question is of some import given the growing international concern about the need to encourage more students to pursue the further study of science to sustain the scientific and technological base of advanced societies (Lord Sainsbury of Turville, 2007; National Academies of Sciences, 2010; Tytler et al., 2008).

## Summary and Research Questions

Therefore, to test these ideas, and specifically the effect of this intervention on student outcomes, this project sought to engage in a minimalist program of professional development and support with four school science departments over a period of 2 years to investigate the following four questions:

(1) Does a minimalist program of professional development designed to encourage teachers to engage students in argumentation on a regular basis lead to an improvement in their students':
   (a) Ability to reason?
   (b) Their knowledge of the content of science?
   (c) Understanding of the epistemic nature of science?
   (d) Engagement with school science?

Naturally as the intervention was based in a theory of action which sought to change teacher practices, we also investigated what effect this intervention had on the teacher practices. The data and the findings answering this goal will be reported in later papers.

## Methods

To undertake this work, a number of schools in a large urban conurbation in the UK were approached. The four schools chosen by the research team were selected to maximize the range of ability and the socio-economic diversity of the sample. Table 1 provides a summary of the principal details of the schools and their student intake for both the experimental and comparison sample.

The data for these schools reflect a range of diversity around the national mean with the exception of the percentage of students who have official statements that they require special educational provision where all the schools are above the national average. This anomaly is most likely attributable to the fact that all schools were situated in urban environments either with considerable ethnic diversity and a large number of English language learners, and/or immigrant communities with higher than normal numbers of transitory students. For the purposes of research, a comparison sample of schools recruited from similar areas was gathered by recruiting an additional four schools.

### *The Nature of the Intervention*

The work in each school was lead by two "lead teachers." Their role was to attend the 5 days of professional development that was offered across the course of the two school years 2007–2009; to work with their colleagues to embed argumentation activities into their schemes of work for both KS3 (Years 7–9[3]) and KS4 (Years 10 and 11); to support and train their colleagues in the use of argumentation-based, instructional practices; and supply their colleagues with relevant materials. As such, they had a leadership role advising their colleagues about appropriate instructional practices and organizing periodic meetings (which they were advised to hold once a month) to reflect on what was being learnt from the use of such activities. These teachers were willing volunteers but not necessarily knowledgeable about argumentation. In one school, both lead teachers left after 1 year and their place was taken by two alternate teachers. The number of teachers participating meaningfully in this program in any one school varied from four to nine teachers and our data would suggest that they had varying levels of commitment to the objectives of the program. Further details of the practice of these teachers, how they changed and the way in which these teachers worked with their colleagues will be reported later.

The professional development days were led by the researchers and made use of the video based professional development materials produced for the IDEAS (Ideas, Evidence &

Table 1
*The major characteristics of schools in the experimental sample*

| School | Type | % Free School Meals | % School Obtaining 5 + A–C Grades at GCSE Including English and Maths (2008) | Persistent Absence Rate | Percentage of Students with Special Educational Needs (Statemented) |
|---|---|---|---|---|---|
| Westside | Comprehensive Mixed | 11% | 66% | 3.5% | 5.8% |
| Hamside | Comprehensive Girls | 23% | 79% | 0.5% | 4.5% |
| Eastside | Comprehensive Mixed | 27% | 40% | 9.5% | 8.3% |
| Northside | Comprehensive Mixed | 33% | 38% | 5.2% | 17.7 |
| National average | | 15.4% | 48.2 | 6.4 | 2.0% |

Argument in Science Education) project (Osborne et al., 2004b). These materials address six themes which previous work with teachers had found to be important in establishing a more argument-based approach to teaching science. The first theme explores what is meant by argument, why it is an important feature of science lessons and how it can lead to enhanced student learning. The second theme explores the pragmatic issue of how to organize and scaffold student discussions so that they are productive. Research to date would suggest that their use for anything other than laboratory work is uncommon for most teachers of science (Newton, Driver, & Osborne, 1999; Weiss et al., 2003). The third and fourth theme explore a range of ways of establishing a context for argumentation in the classroom using strategies such as listening triads, argument lines, jigsaw groups, and some of the resources that are available to support the teaching of science using argumentation. Theme 5 of evaluating argument uses a set of video-based exercises to help build teachers understanding of the features of an argument that distinguish weak and good arguments. The final theme explores how teachers can model what constitutes an argument in science to communicate the kind of practice that students might engage in. The professional development with the lead teachers drew extensively on these materials to help build their conception of the goal of the intervention, how it could be taught and what to communicate to their colleagues. In addition, some time was devoted to exploring their current schemes of work and how these could be adapted to include specific lessons which incorporated argumentation on a regular basis. Plans and ideas were then shared and discussed by the teachers. At each workshop, the lead teachers shared these experiences before experiencing further inputs. The programme provided opportunities for the teachers to share resources and instructional strategies that could be taken back to their science departments and shared with their colleagues. Time was also given over to reflective analysis and consideration of how to work with their colleagues effectively.

Between workshops, the research team visited schools on a regular basis to collect data from teacher interviews, observations, and reflective meetings. Though some discussion between researchers and teachers took place after observations of lessons, prolonged systematic feedback was not provided for two reasons. First the aim was to research the process of change through a light touch from the team, relying more on the processes initiated by the lead teachers within the department. Second, consistency of feedback across all teachers and schools would not have been possible. Rather, drawing on the Clarke and Hollingsworth, interconnected model of professional growth the intervention relied upon the lead teachers to initiate this cycle of professional growth by providing a stimulus for promoting the use of argument and professional experimentation, and on the reflective meetings for identifying and reviewing salient outcomes, changing teacher knowledge and beliefs, and developing practice.

Commonly, as a stimulus for professional experimentation in each school, argumentation lessons were developed and tested by individual teachers and then embedded in the schemes of work which other teachers were expected to use. Reflective meeting then provided a space for distributed leadership, where all teachers, both leaders and followers, could consider how these practices functioned in their schools (Spillane, 2006). For instance, in one such meeting, the lead teacher chaired a meeting held specifically to discuss Talking to Learn. This started with each teacher sharing a descriptive/structural account of one argumentation activity they had used. These descriptive accounts triggered some comparisons and discussion of rationale with other teachers. A more general discussion followed, focusing on group work, how pupils were developing argumentation skills, and how teachers had developed their practice.

A number of issues emerged in the reflective meetings. Argumentation was perceived to be difficult for less able students to undertake. Some students had difficulty expressing arguments and teachers ascribed this difference to a lack of vocabulary or a lack of conceptual understanding in expressing opinions. Some teachers also felt that argumentation lessons that addressed both science content and epistemic aspects of science could be confusing for students. Teachers reported a lack of confidence when teaching what they perceived to be less "controlled and calm and organized" lessons, including behavior management issues. They also found it difficult to stimulate disagreement and discussion as students tended to agree with each other too often. Other teachers were concerned about how frequently to use argumentation lessons in the curriculum, or how to use 50 minutes of the lesson to best effect without running out of time. Teachers were also focused on how to group pupils so that all students worked well, and the balance between writing and talking within argumentation lesson. This balance was expressed as a tension between wanting to maximize discussion and thinking versus the value of writing as a means of expressing and recording student thinking. Schools were asked to hold reflective meetings once a month to discuss and explore progress but the exigencies and imperatives of schools meant that such meetings were held less frequently than this varying by school between 4 and 8 meetings per year.

The range of issues raised demonstrates the complexity of the professional learning process in this study, which, in spite of previous studies, could not be fully anticipated. Though the lead teachers were all invited to the five teacher meetings, in some instances they did not all attend, so there was variation in how they experienced the inputs. In addition, the range of confidence and experience in science teaching, argumentation practice and leadership was greater than anticipated, and reflective meetings were fewer than anticipated.

Thus, there was considerable variation in the way that both the lead teachers and their colleagues adopted and used argumentation-based activities in their classrooms. In part, this is because what the teachers were asked to implement was a set of instructional practices and not any specific curriculum materials. The latter would have been impossible given the diverse nature of the grades the teachers were working with and the range of curricula offered by each school. And in part also, this was a deliberate choice as the notion that even the most highly scripted curriculum can ever be implemented with a universal consistency of treatment is a chimera. Rather teacher learning should recognize that adaptation and customization are inherent features of learning (O'Donnell, 2008; Ogborn, 2002). No two teachers are alike and no two students will respond to any stimulus in similar manners. By acknowledging the complexity of working with teachers in highly variable contexts (Opfer & Pedder, 2011), we sought to see if there were any identifiable differences in outcome between the intervention and comparison schools. Our concern here was not with fidelity of implementation in the classroom which, given the variation in contexts, curricula, and schemes of work would have been impossible. Rather, it was with teachers taking ownership and adoption of a style of teaching that might make a measurable difference to student outcomes—measurements of which are the focus of this study.

### *Rationale for Instruments*

*Scientific Reasoning and Conceptual Knowledge Gains.* To answer our questions, extensive consideration was given to the nature of the instruments to be used for measurement. Argumentation does require students to engage in higher order cognitive demands of synthesizing an argument, juxtaposing alternatives, and evaluating their competing merits. Valid tests of reasoning tend to be restricted either to highly specific aspects of scientific reasoning such as the ability to control variables found in the kinds of tasks developed by Shayer, Wylam, Adey, and Kuchemann (1979), each of which takes considerable time to administer, or to domain-general aspects of reasoning such as the Cornell test of critical thinking (Ennis, Millman, & Tomko, 1985). The focus of the latter type of tests, however, is on measuring students' ability to use deductive logic rather than on students' ability to use informal reasoning and argumentation.

In their work on the use of exploratory talk, Mercer and his colleagues had demonstrated that the use of explicit, collaborative reasoning had led to significant gains in scores on Raven's Progressive Tests of non-verbal reasoning (Mercer, Wegerif, & Dawes, 1999). This finding had been replicated in a later study with an explicit focus on learning science (Mercer et al., 2004). Raven's matrices have also been validated by many replications and are also normalized against the cohort of representative populations in different countries for different ages (Raven, Raven, & Court, 2004). As a test, Raven's matrices are also relatively straightforward to administer and the measurement is not cofounded by students' linguistic abilities. Hence, the English language learner should be able to perform on this test as well as a native speaker of similar cognitive ability. Thus, whilst not a direct measure of students' ability to reason argumentatively in the context of science, for reasons of validity and pragmatism, this instrument was used to collect data relevant to Q1a. Data from the normalization studies for the UK population of this age showed that Both the Year 7 cohort and the Year 9 cohort were normally distributed but the Year 7 cohort was skewed towards the upper quartile of the population, particularly for the experimental group. Likewise Year 9 cohort were also skewed toward the upper quartile though equally for both cohorts.

To answer question Q1b, one option would have been to use a standard test of conceptual knowledge. However, given the diversity of syllabi and experiences, no single test exists which would has universal validity. For these reasons, it was decided to use the data from national standard tests of student knowledge of science. These took two forms—data available from tests that students took at age 11 and at age 14 (known as KS2 and KS3 tests), and data available from their performance in national examinations at age 16. Two caveats must be applied to this data. First, the validity and reliability of these data sources is open to question as no data are published on how these measures are established. Second, the examination that students take at age 16, whilst conforming to national requirements, are not all the same examination though mechanisms do exist for moderation to ensure the equivalency of the grades.

*Students' Understanding of Scientific Epistemology and Their Epistemic Beliefs.* What features of student understanding of the nature of science might be affected by engaging in argumentation was the focus of Q1c. Several aspects emerged for consideration. First, engaging in argumentation requires an exploration of the relationship between theory and evidence and the ability to discriminate relevant from irrelevant evidence. Second, examining the relationship between theory and evidence might lead to a change in students' epistemic beliefs and an improvement in their view and understanding of the role and relative importance of theories and facts, if only tacitly. In short an improvement in what might be seen as epistemic knowledge. Third, engaging in argumentation requires the evaluation of the merits of different data and their associated warrants which might lead students to take a more evaluativist stance to knowledge and be reflected in students' personal epistemologies.

To measure students' epistemic beliefs, we drew on the work of Hofer and Pintrich (1997) who have suggested that there are four dimensions to any individual's epistemological beliefs. These are: certainty of knowledge (stability); simplicity of knowledge (structure); source of knowing (authority); and justification for knowing (evaluation of knowledge claims). *Certainty of knowledge* is something which changes over time beginning at lower levels where knowledge is perceived to be certain and absolute and developing to a view where knowledge is seen as tentative and evolving and where individuals are open to new interpretations. *Simplicity of knowledge* is a measure of an individual's understanding of the structure of knowledge and lies on a continuum between seeing knowledge as an accumulation of facts versus a set of highly inter-related concepts. *Source of knowing* refers to views an individual holds about the source of epistemic authority. At lower levels it originates outside the self and resides in an external authority. A substantial transformation occurs when an individual who was previously "a holder of meaning becomes a maker of meaning" (Perry, 1970) (p. 87) and recognizes that transformation. *Justification of knowing* refers to the views that individuals hold about how claims to knowledge can be justified moving from dualistic beliefs where ideas are either right or wrong to an understanding that knowledge is something which requires a reasoned justification for belief. To measure these concepts, an instrument was initially developed by Elder (2002) and then improved by Conley, Pintrich, Vekiri, and Harrison (2004), who showed it to be internally reliable and hold a four-dimensional factor structure corresponding to these theoretical scales using confirmatory factor analysis for testing Grades 4–5 students. While our students were Grades 6–8 students, it was not felt that the small age difference would affect the validity of the instrument. Hence, it was the instrument we chose to use in this research. Analysis of our data using exploratory factor analysis also showed that the data fitted a four factor model with each of the scales having a Cronbach alpha >0.75.

When it comes to assessing students' personal (or practical) epistemologies, Kuhn and her co-workers (Kuhn, 1999; Kuhn, Cheney, & Weinstock, 2000) have argued for a developmental model proposing that students begin life as "absolutists" where knowledge consists of certain and unequivocal facts. From here, students are thought to progress to becoming essentially relativists. This transition occurs when they begin to recognize that there is a subjective element to knowledge. However, they still lack a sense that there are criteria that enable some claims to be judged more valid than others. These individuals Kuhn et al. term as "multiplists." The final stage is when students become "evaluativists" and recognize that there is a need to evaluate critically the evidence used to support claims to knowledge. To assess the nature of students' personal epistemologies we used three items developed by Kuhn and her co-workers which were combined with the data from the Conley instrument. These essentially provided an alternative means of measuring students' epistemic beliefs though the number of items is too small to make this element a reliable measure. Combining both measures into a single survey instrument resulted in a questionnaire consisting of 38 Likert items.

*Engagement With and Attitudes Towards Science.* Measurement of attitudes towards science, the focus of question Q1d, has been a field which has been beset by issues of validity and reliability (Blalock et al., 2008). To measure student attitudes therefore we drew on two instruments for which factor analysis had demonstrated independence of the constructs and good measures of internal reliability, that is, Cronbach alpha >0.7. One instrument was developed by Nolen (2003) to measure learning environment, motivation, and achievement in high school, and one developed by Pell and Jarvis (2001) to measure students' attitudes to science for Grades 4–5 students. Given that both of these instruments had been designed and tested with students who only differed in age by approximately 2 years from our students, it was felt that they would still provide valid measures of both engagement and attitudes towards science. From the instrument used by Nolen we drew on the items used to measure the extent to which individuals views were dominated by the importance of their success (ego orientation) versus the extent to which the task itself was intrinsically worthwhile (task orientation).

In addition we used items which measured the extent to which students sought to avoid working (work avoidance), were satisfied with their learning (satisfaction with learning), perceived the classroom to have a cooperative and collaborative climate (collaborative climate) or an emphasis on producing the right answer (ability and meritocracy). Finally, we used a set of items from both Nolen and Pell and Jarvis to assess interest in science. Both instruments were used to improve the reliability as Pell and Jarvis' instrument measured attitudes towards science in general rather than just school science. This resulted in a questionnaire consisting of 74 Likert items which was administered using a computer or, where preferred, on paper. This instrument can be found in the Supporting Information. An initial analysis of our data set confirmed that each of the scales had a Cronbach alpha >0.76.

## Results and Findings

For the purposes of answering questions 1a, 1c, and 1d, each school was asked to identify two classes in Year 7 and two classes in Year 9 from which to collect data. Thus approximately 60 students in each year from each school were tested for their ability to reason, their attitudes towards science, and their understanding of the nature of science in both the intervention and the comparison schools, using the instruments described previously. Students for testing were selected by the teachers in consultation with the researchers who asked for classes that were representative of a cross-section of students in the school. The Year 7 students are between 11 and 12 years old and were 58% girls, 42% boys (comparison schools) and 66% girls, 34% boys in the experimental schools. Year 9 students are 13–14 years old and were 31% boys, 69% girls in the comparison schools and 26% boys and 74% girls in the experimental schools. All of these instruments were initially taken by all of the experimental and comparison students in the autumn of 2007. The instruments were then retaken, 18 months later between March and June in 2009. The Raven's matrices were administered as pencil and paper tests and all of the other instruments using a computer-based survey in class time. These students were then taught in some cases by the lead teachers for the next 2 years and in some cases by other teachers in the department. Tracking which student was taught by which teacher was not possible. Exact sample sizes varied depending upon the number present for any given test but varied between a maximum of 391 for the Raven's Matrices Test and a minimum of 292 for the assessment of their epistemic beliefs for the experimental schools. Comparable figures for the comparison schools were a maximum of 354 and a minimum of 315. Data for answering question 1b—how students conceptual understanding had changed—was obtained from the schools who provided the information on how their students had performed on national tests.

The Ravens' Matrices require students to select one response from a choice of 6. The survey items used a 5-point Likert-type scale. Inevitably some responses were missing because students had failed to complete all items. Where this happened, the responses were examined and those which were incomplete at the end of the test were deleted as these indicated test fatigue and it was not possible to infer what responses might be. For the other responses *t*-tests were run that showed that the data were not missing completely at random. Hence, the SPSS package which imputes missing data using the expectation–maximization algorithm was then applied to the data.

### *Measures of Non-Verbal Reasoning Ability (Raven's Matrices)*

The outcome for the RM data was 745 responses for pre- and post-data. The RM test has a score which ranges from a 0 to 49. Table 2 shows the mean scores and standard deviations prior to the intervention.

Table 2
*Pre-intervention scores on Raven's matrices for control and experimental schools*

| Schools | Year Group | Mean | *SD* | *N* |
|---|---|---|---|---|
| Experimental | 7 | 33.67 | 5.32 | 201 |
| | 9 | 35.21 | 5.42 | 165 |
| | Total | 34.36 | 5.41 | 366 |
| Control | 7 | 31.81 | 5.54 | 200 |
| | 9 | 36.01 | 5.11 | 179 |
| | Total | 33.79 | 5.73 | 379 |

Analysis of the means showed that there was a significant difference between the Year 7 group (age 11) and the Year 9 cohort (age 14) as might be expected given that Raven's Matrices is a measure of reasoning which has been shown to develop with age. Two way ANOVA's conducted on the Year 7 cohort showed that there was a significant difference prior to the intervention ($F$(1, 400), $p = 0.001$) but there was no significant difference for the Year 9 group $F$ (1, 343), $p = 0.163$.

Table 3 shows the mean scores and standard deviations obtained after the intervention. An analysis of co-variance (ANCOVA) of the post-intervention scores using the pre-intervention score as the co-variate showed that there was a significant difference between the experimental and comparison groups ($F$(1, 743) = 25.928, $p < 0.001$) with the comparison group having a higher adjusted mean score after the intervention. An ANCOVA conducted on the Year 7 (age 11) cohort showed that there was a significant difference after the intervention ($F$(1, 400) = , $p = 0.01$) with the comparison group having the higher adjusted mean score. Likewise there was a significant difference for the Year 9 (age 14) group $F$(1, 343) = 16.968, $p < 0.001$. with the comparison group having a higher adjusted mean score after the intervention. Full descriptions of all the ANCOVA tests are provided in the Supporting Information.

A Tukey post hoc analysis of the score gains on this test for Year 9 students indicates that the gains were not homogeneous. One set of schools were the mean gain ranged from −1.269 to +.7632 contains three comparison schools and one experimental school (Group A). These differed significantly ($p < 0.05$) from Group B (one experimental and one comparison school) with an average mean gain = +1.5. Both Group A and Group B differed significantly ($p < 0.05$) from one experimental school for whom the mean gain was +4.24.

Likewise, the Tukey post hoc analysis for Year 7 students shows that the gains were also not homogeneous. Two comparison schools (Group C) with an average mean gain = +0.55 differed significantly ($p < 0.05$) from Group D (to comparison schools and two experimental

Table 3
*Post-intervention scores on Raven's matrices for control and experimental schools*

| Schools | Year Group | Mean | *SD* | *N* |
|---|---|---|---|---|
| Experimental | 7 | 34.82 | 5.03 | 201 |
| | 9 | 35.43 | 6.27 | 165 |
| | Total | 35.02 | 5.62 | 366 |
| Control | 7 | 34.69 | 5.65 | 200 |
| | 9 | 38.17 | 6.11 | 179 |
| | Total | 36.34 | 6.12 | 379 |

Table 4
*Pre-intervention scores on KS3 tests for control and experimental schools*

| Schools | Year Group | Mean | *SD* | *N* |
|---|---|---|---|---|
| Experimental | 7 | 4.61 | 0.508 | 205 |
| | 9 | 5.74 | 0.982 | 193 |
| Control | 7 | 4.21 | 0.888 | 146 |
| | 9 | 5.87 | 0.933 | 196 |

schools with an average mean gain of 1.89. Group D with the addition of one more experimental school then differs significantly ($p < 0.05$) from one experimental school for whom the mean gain was +4.5. Taken together this suggests that there was no consistent effect for either the experimental or the comparison group.

*Knowledge of Science*

The measure of student understanding of science content knowledge prior to the intervention was the level attained by each student in the national tests in 2007/2008. This was measured by two standard national tests, one for students of age 10–11 (KS2) and another for students age 13–14 (KS3), both of which are graded externally. On these tests, students are graded in terms of a metric which ranges from 1 to 7 in unit intervals which is applied consistently across both tests. For the Year 9 cohort, the measure of their understanding of science content at the end of the intervention was their mean performance in the national examinations which are awarded a Grade $A^*$ to G (recoded $A^* = 8$ to $G = 1$). For the younger cohort it was their performance on the KS3 test. For the younger cohort, policy changes between starting and finishing the intervention meant that the KS3 tests were no longer administered nationally. Rather, schools have taken to using old tests to obtain a value for student performance and grading the results themselves.

Cases where data were missing either before or after were deleted. Table 4 shows the mean scores and standard deviations prior to the intervention.

Using a two-way ANOVA showed that there was no significant difference in the student performance scores before the intervention between the Year 7 (age 11) cohorts in the experimental and comparison schools ($F(1, 388)$, $p = 0.176$). There was, however, a significant difference between the Year 9 (age 14) groups in the experimental and comparison schools ($F(1, 350)$, $p < 0.001$) before the intervention with the experimental cohort performing significantly better.

Table 5 shows the mean scores and standard deviations obtained after the intervention. In this case, because of the categorical nature of the data, gain scores were calculated and a comparison made of the gains which showed that there was no significant difference between the experimental and comparison schools for the Year 7 (age 11) cohort with chi-square (5, 336) = 0.118). The same analysis for the Year 9 (age 14) cohort showed that there was a

Table 5
*Post-intervention scores on YEAR 7/GCSE tests for control and experimental schools*

| Schools | Year Group | Mean | *SD* | *N* |
|---|---|---|---|---|
| Experimental | 7 | 5.88 | 0.866 | 197 |
| | 9 | 5.55 | 1.16 | 185 |
| Control | 7 | 5.61 | 1.067 | 139 |
| | 9 | 5.93 | 1.359 | 191 |

significant difference after to the intervention (28, 375), $p = 0.002$) with the comparison group having made a larger gain.

### *Measures of Student Understanding of Scientific Epistemology and Their Epistemic Beliefs*

Student understanding of scientific epistemology was measured using the scales developed by Conley for their views about source of knowledge, the certainty of knowledge, the development of knowledge and the justification of knowledge. These scales use Likert items on a 1–5 categorical scale. The higher the number, the more sophisticated the measure of student understanding. Table 6 beneath shows the means for the Year 7 (age 11) sample. Once again, an analysis of the missing data was conducted. Those where whole data sets were missing were deleted. For the rest (<5% of the sample), the data were imputed using all the expectation-maximization algorithm of SPSS and all the variables.

Two way ANOVAs show that there was a significant difference between the means for the comparison and experimental group for the scales testing source of knowledge ($F(1, 346)$, $p = 0.002$) (experimental higher), and the certainty of knowledge ($F(1, 346)$, $p < 0.001$) (experimental higher) but no difference for the development of knowledge ($F(1,346)$, $p = 0.525$) and the justification of knowledge ($F(1, 346)$, $p = 0.455$). The relatively high scores on these last two scales may be indicative of a saturation effect. The data for the Year 9 group are shown in Table 7. None of the differences for this group were found to be significant.

The results for the Year 7 cohort after the intervention are shown in Table 8. An ANCOVA analysis of these scores shows that there were no significant differences between the experimental and the comparison group.

An ANCOVA analysis for the Year 9 cohort is shown in Table 9. Here, there are significant differences on three scales: Source of knowledge ($F(1,266) = 5.09$, $p < 0.05$) where the comparison group have the significantly higher adjusted mean score after the intervention; development of knowledge ($F(1,266) = 5.387$, $p < 0.05$ where the comparison have advanced more; and the justification of knowledge ($F(1, 266) = 4.529$, $p < 0.05$) where the experimental group have advanced more. However, many of the differences are not large and given the size of the sample it would be a mistake to attribute much significance to them especially as the effect sizes are small (partial eta square between 0.017 and 0.022).

Table 6
*Pre-intervention scores on measures of student understanding of scientific epistemology for Year 7 cohort*

| Beliefs About | Control or Experiment | Mean | *SD* | *N* |
|---|---|---|---|---|
| Source of knowledge | Expt | 3.51 | 0.759 | 171 |
| | Control | 3.24 | 0.840 | 176 |
| | Total | 3.37 | 0.811 | 347 |
| Certainty of knowledge | Expt | 3.65 | 0.647 | 171 |
| | Control | 3.36 | 0.788 | 176 |
| | Total | 3.50 | 0.735 | 347 |
| Development of knowledge | Expt | 4.17 | 0.385 | 171 |
| | Control | 4.14 | 0.456 | 176 |
| | Total | 4.16 | 0.422 | 347 |
| Justification of knowledge | Expt | 4.22 | 0.372 | 171 |
| | Control | 4.19 | 0.413 | 176 |
| | Total | 4.20 | 0.393 | 347 |

Table 7
*Pre-intervention scores on measures of student understanding of scientific epistemology for Year 9 cohort*

| Beliefs About | Control or Experiment | Mean | *SD* | *N* |
|---|---|---|---|---|
| Source of knowledge | Expt | 3.66 | 0.747 | 100 |
| | Control | 3.66 | 0.679 | 168 |
| Certainty of knowledge | Expt | 3.82 | 0.593 | 100 |
| | Control | 3.70 | 0.619 | 168 |
| Development of knowledge | Expt | 4.17 | 0.395 | 100 |
| | Control | 4.18 | 0.408 | 168 |
| Justification of knowledge | Expt | 4.14 | 0.351 | 100 |
| | Control | 4.12 | 0.4107 | 1168 |

Table 8
*Post-intervention scores on measures of student understanding of scientific epistemology for Year 7 cohort*

| Beliefs About | Control or Experiment | Mean | *SD* | *N* |
|---|---|---|---|---|
| Source of knowledge | Expt | 3.59 | 0.748 | 100 |
| | Control | 3.77 | 0.689 | 168 |
| Certainty of knowledge | Expt | 3.84 | 0.631 | 100 |
| | Control | 3.89 | 0.632 | 168 |
| Development of knowledge | Expt | 4.12 | 0.486 | 100 |
| | Control | 4.28 | 0.586 | 168 |
| Justification of knowledge | Expt | 4.06 | 0.467 | 100 |
| | Control | 4.17 | 0.452 | 168 |

The analysis of the data for students' epistemic beliefs about the nature of knowledge using the Kuhn items is shown next. Because of the constraints of time the survey could only use three items to measure students' beliefs and, given that, the reliability is in doubt. The data are presented here, nevertheless, because such large scale measurements are few in the literature and the outcomes are somewhat surprising. Kuhn's items divide students into absolutists (those that believe knowledge is absolute and certain), multiplists (those who believe

Table 9
*Post-intervention scores on measures of student understanding of scientific epistemology for Year 9 cohort*

| Beliefs About | Control or Experiment | Mean | *SD* | *N* |
|---|---|---|---|---|
| Source of knowledge | Expt | 3.63 | 0.708 | 171 |
| | Control | 3.58 | 0.801 | 176 |
| Certainty of knowledge | Expt | 3.83 | 0.646 | 171 |
| | Control | 3.69 | 0.758 | 176 |
| Development of knowledge | Expt | 4.22 | 0.537 | 171 |
| | Control | 4.13 | 0.419 | 176 |
| Justification of knowledge | Expt | 4.22 | 0.484 | 171 |
| | Control | 4.14 | 0.464 | 176 |

Table 10
*Students' personal epistemological beliefs: Percentage of students holding each category*

| Year | Absolutists | Multiplists | Evaluativists | Mixed |
|---|---|---|---|---|
| Pre | | | | |
| Year 7 | | | | |
| Expt ($n = 169$) | 1.8 | 14.2 | 76.3 | 7.7 |
| Control ($n = 167$) | 15.0 | 14.4 | 64.1 | 6.6 |
| Year 9 | | | | |
| Expt ($n = 94$) | 5.3 | 17.0 | 71.3 | 6.4 |
| Control ($n = 168$) | 7.1 | 13.1 | 75.6 | 4.2 |
| Post | | | | |
| Year 7 | | | | |
| Expt | 4.7 | 11.8 | 77.5 | 5.9 |
| Control | 7.8 | 13.2 | 74.3 | 4.8 |
| Year 9 | | | | |
| Expt | 13.8 | 13.8 | 69.1 | 3.2 |
| Control | 6.0 | 7.7 | 83.9 | 2.4 |

that there are no criteria to distinguish between claims to knowledge) and evaluativists (those who believe that claims to knowledge can be evaluated against agreed criteria). Table 10 shows the distribution across the whole sample for each Year and includes a column of "mixed" where the students adopted a different position on each question.

The main feature of note to emerge from this analysis is the large number of students who seem to be operating with an evaluativist epistemology—a stance which is commonly regarded as the most sophisticated form of epistemology. This finding contradicts that of Kuhn et al. (2000) who found that only 32% of eighth graders and 38% of 12th graders held such a stance. A chi-square test shows that the distributions were significantly different between the experimental and the comparison group in the Year 7 ($p < 0.001$) cohort prior to the intervention but not after. This is due predominantly to the difference in the number of absolutist thinkers (more in the comparison group) initially which had disappeared after the intervention. For the Year 9 group, there was no significant difference beforehand but there was afterward ($p < 0.05$) which was due to the enhanced number of absolutists in the experimental group afterwards and the enhanced number of evaluativists in the comparison group. None of these effects are large compared to the predominance of evaluativists overall and suggesting that the intervention has not had a substantial effect on students' personal epistemological beliefs.

### *Measurement of Student Attitudes and Engagement*

The attitude instrument consisted of 74 items which measured 11 constructs on independent scales which was confirmed by a factor analysis. In addition, a 12th single item tested the extent to which students perceived discussion to be a feature of the classroom environment. A two-way ANOVA was conducted to identify significant differences that might exist between the samples beforehand and then an ANCOVA on the post-test results using the pretest response as the covariate. Significant differences identified are shown in Table 11 with an indication of which group has improved more.

The general features are that prior to the intervention the experimental group in Year 7 had significantly better scores on the constructs of task orientation (the extent to which they viewed learning as an end in itself), the work aversion scale (indicating a greater willingness

Table 11
*Summary of data for attitude and engagement constructs*

| Construct | Year 7 Pre | Year 7 Post | Year 9 Pre | Year 9 Post |
|---|---|---|---|---|
| Ego orientation | | Control improved more ($p = 0.045$) | Control Better ($p = 0.016$) | |
| Task orientation | Experimental better ($p = 0.01$) | | Control Better ($p = 0.025$) | |
| Work aversion | Experimental better ($p = 0.025$) | | | |
| Satisfaction with learning | | | | [a] |
| Collaborative climate | | Control improved more ($p = 0.022$) | | |
| Science learning focus | | | | [a] |
| Ability and meritocracy | | | | |
| Interest in learning science | Expt better ($p = 0.003$) | | Expt better ($p = 0.011$) | Expt improved more ($p = 0.001$) |
| The social value of science | Expt better ($p < 0.001$) | Experimental improved more ($p < 0.05$) | Expt better ($p = 0.01$) | |
| Science enjoyment | Expt better ($p < 0.001$) | | | |
| Difficulty in science | | | | Control improved more ($p = 0.032$) (not so difficult) |
| Discussion in science | | Experimental improved more ($p < 0.05$) | | |

The following table summarizes the information obtained from all of the tests. Full details and statistics are available in the Supporting Information.

[a] Conditions for ANCOVA test were not satisfied. See the Supporting Information.

to undertake the work required), their interest in learning science, the social value of science, and their enjoyment of learning science. However, after the intervention, these differences had disappeared and the comparison group had improved more on their perception that there was a collaborative climate in the classroom and ego orientation (indicating that they were less focused on their own achievement).

The picture for Year 9 is more mixed with the comparison group initially having a more positive ego and task orientation that had disappeared 18 months later. Conversely, the experimental group had more interest in learning science and placed a higher rating on the social value of science. Eighteen months later, interest in learning science was even stronger for the experimental group but the difference in their ratings of the social value of science had disappeared. The other change was that the perceived difficulty of science had reduced more for the comparison group than the experimental group.

Thus, it would seem that the intervention has had some negative effects on interest and engagement for the experimental group in Year 7 whilst the picture for the Year 9 cohort is that overall it has made little change apart from the one notable effect of improving their interest in learning science.

## Discussion

Overall, the picture obtained from this research then is essentially one of null findings—that is the intervention has not shown any consistent effects for student reasoning, student engagement, or an improvement in their epistemic understanding. Nevertheless, we would still argue that there is much to be learnt from their publication. First, such findings establish a deeper understanding of what is, and what is not achievable by professional development. Findings such as these then act to prevent the bias that arises in our understanding of a phenomenon (Lehrer, 2010; Moller & Jennions, 2001) such as that identified by Thomas Sterling who noted in 1959 that 97% of all published studies in psychology found the effect they were looking for with statistical significance (Sterling, 1959).

Any educational intervention must compete for success in the maelstrom that constitutes the daily life of the teacher. This work sought to see whether such practices could gain acceptance and a toehold within the common practice of the teaching of school science *without* extensive professional development and coaching and affect student outcomes. Whilst evidence exists that professional development can be successful at significantly enhancing student outcomes (Yoon et al., 2007), what has been shown to work to date is costly and many school systems cannot afford to support teacher professional development to this extent.

Thus, this research was an attempt to see if change could be effected with a minimal intervention. The school case studies, yet to be reported, demonstrate that changes in curriculum content and teaching approaches took place in all four schools, but the extent and nature of change was variable. One detailed case study of two of the teachers conducted by Christodoulou (2011) suggests that the teachers' use of argumentation was still developing 18 months after the start of the project. In addition, the complexity of school policy, departmental culture and leadership style, together with differences in the expertise, beliefs, priorities, and teaching goals of the individual teachers, resulted in four quite different pictures of innovation.

Moreover, while all the schools volunteered to participate in the project, the individual teachers themselves were not necessarily volunteers. Clearly, it simplifies the testing of any scheme to work solely with teachers who are committed to the educational values and perspectives of any project. However, testing any kind of intervention at scale must demand the participation of teachers who have been *required* to use a new approach. Only then would the trial of the innovation be a true test of what the outcomes might be when implemented at scale. To date, however, work in professional development has used groups of volunteer teachers. Little is known how the findings might differ if studies were not limited to motivated volunteers. Thus, the context for this study is a more accurate reflection of the range of commitments that might commonly be found in schools for learning new strategies and practices.

In their review of what works in professional development, Guskey and Yoon (2009) point out that none of the interventions that showed positive effects had <30 contact hours of professional development and none of these used a train the trainer approach. Given that <30 hours contact time of professional development was offered to all of the teachers other than the lead teachers and that the model here was one of training the trainers, these findings give further support to the conclusion reached by Guskey and Yoon and suggest that attempts to circumvent this amount of training such will not succeed. Indeed, duration is one of the five core features of successful professional development identified by Desimone (2009). The other four are content focus, active learning, collective participation, and coherence with teachers' knowledge and beliefs. While the program contained four of these features, the extent to which it was coherent with the knowledge and beliefs of the teachers did vary from school to school and from teacher to teacher. Moreover, the framework we were using was based on the idea that teacher development would be essentially internal to each school and teacher driven. In their work testing the STELLA model of professional development, Roth et al. (2011) found such an approach lacked sufficient focus to develop the specific change they sought. Hence, they imposed a more externally driven framework to maximize their impact. Thus, a primary reason for the null findings of this study may be that this kind of intervention does not change teacher practice sufficiently to affect student outcomes.

Another possible reason for the null effect we obtained is that our instruments did not measure the variables that we sought to measure—that is students' facility with reasoning, their engagement with science, or the changes in their epistemic beliefs. This argument is possibly most valid for Raven's matrices as a test of reasoning as these are non-verbal and domain free. Given the argument that reasoning is domain dependent and that scientific reasoning requires scientific knowledge (Perkins & Salomon, 1989), there is some justification to this point. The counter-argument is that argumentation requires the discursive use of evaluation which sits at the top of a cognitive hierarchy which goes from recall to explanation to juxtaposition to evaluation itself (Klahr, Zimmerman, & Jirout, 2011). Argumentation therefore requires the performance of evaluation and the exercise of reasoning—the very ability which is measured by Raven's matrices (Billig, 1996; Kuhn, 1992)—albeit non-verbal. Moreover, what we found was no effect on nearly all measures rather than any one. This outcome would suggest that it is not flaws in the instruments itself which might account for the null findings.

Our view is that what the field can learn from this study comes from an analysis of the constraints within which the study was conducted. First, we were asking teachers to adopt a new instructional practice. An accumulating body of empirical evidence (Martin & Hand, 2009; Ratcliffe, Hanley, & Osborne, 2007) would suggest that it takes a minimum of 2 years for teachers to begin to use new instructional practices in an effective manner. The first year requires teachers to attempt new practices that involve risk and uncertainty (Claxton, 1988). Inevitably, teachers will make mistakes raising the issue of whether there has been fidelity of implementation—that is was the practice implemented in a manner which was the intention of the researchers. In the research conducted by Martin and Hand (2009)—a longitudinal case study of one experienced teacher's attempt to embed elements of argumentation into her science classroom—the teacher took 2 years to become competent with the practice. Likewise, Ratcliffe et al. (2007) found that it took a 2-year cycle of one new, innovative, junior high-school curriculum for teachers to become familiar with the goals and the appropriate instructional practices. Similar findings emerge from the work of Adey and Shayer who found no immediate effects on student performance when students were tested immediately after their "Thinking Science intervention" (Adey & Shayer, 1993) and more recently in the work of Allen, Pianta, Gregory, Mikami, and Lun (2011) who found no student effect until the post-intervention year. Looked at from the perspective of the interconnected model of Clark and Hollingsworth (2002), it would suggest that time is required to forge the connections that lead to change between the domain of practice, the domain of consequences, and the personal domain.

Moreover, for teachers, professional learning is not just a case of developing a new skill but also one of developing a deeper understanding of the theoretical rational of any practice. Achieving such an understanding requires a mix of explicit discussion and reflection on practice that varied considerably in the experimental schools. Other empirical evidence would suggest that achieving competence with new practices takes time. Indeed, the work of Roth et al. (2011) on the Science Teachers Learning from Lesson Analysis (STeLLA) project would suggest that a minimum of 100 hours of professional development is required to achieve changes in teachers' instructional practice that leads to enhanced student learning. Even then, as Garet et al. (2011) have shown in their systematic study of an extensive (and expensive) mathematics professional development program for middle school teachers, significant change is not ensured. Heller, Daehler, Wong, Shinohara, and Miratrix (2012) recently published research on professional development using 270 schools and 7,000 students does show positive effects on teacher knowledge and student attainment but all of the measures were of knowledge or justifications and not of teacher practice. In contrast, Greenleaf et al.'s (2011) intervention did attempt to change teacher practice. Using a group-randomized experimental design with 105 volunteer biology teachers in 83 schools approximately half of whom were assigned to the treatment group and half to the comparison group, this intervention sought to develop high schools teachers' facility to make more extensive use of texts in their teaching. Teacher practices and their changes were measured indirectly from sample lesson materials and samples of student work. Their intervention consisted of a 2-year intervention with summer institutes with follow-up sessions within the school year. Teachers worked in school-based teams that were organized in cross-site learning networks. These researchers found positive effects for students in language arts, reading comprehension, and biology on state standardized tests with effect sizes of 0.23, 0.24, and 0.28, respectively. Notable though is that the professional development had a very specific focus and was undertaken with volunteers.

In contrast, our research had no content-specific focus but its goal was the acquisition of a generic pedagogic practice—the use of argumentation and a more dialogic and interactive pedagogy. And just as the acquisition and use of new word takes 7–10 uses of the word (Waring & Takaki, 2003), a recognition of the word's value and a motivation to use the word fluently and correctly, so does the acquisition of any new practice require an understanding of its value, a motivation to use it, and finite number of attempts to acquire fluency with its use. Thus, the *sine qua non* of professional development has to be first building an understanding for teachers in well-articulated and clear terms of the underlying theoretical rationale and the empirical evidence to justify the value of acquiring the new practice or practices. In this case, that engaging students in argumentation, deliberative discussion and productive talk can change the classroom dynamic to one which is more challenging, stimulating and engaging. To this end, video exemplars of "authentic"[4] classrooms such as those in the IDEAs pack or those offered by the TERC Inquiry Project (http://inquiryproject.terc.edu/) that offer a clear goal of the outcome of competent practice are invaluable. In essence, they provide models of expertise and currently there are insufficient exemplars to use either in initial training or further professional development.

Models of expertise, however, are only a necessary but not sufficient condition for teacher learning as acquisition of a new practice depends on systematic opportunities to build fluency with the practice. Moreover, initial attempts are most likely to fail—just as they would for anybody attempting to hit a golf ball correctly, play a musical instrument, or paint a landscape. In all other domains, competence is acquired by engaging in highly constrained and well-defined exercises which, whilst challenging, are designed to ensure sufficient success to build a sense of progress. In music, for instance, the novice learns how to play simple scales before moving onto exercises involving chords. What we lack currently is a sense of what kinds of initial exercises are needed to develop and implement a more dialogic approach to teaching science, which of these are more crucial, and what is the appropriate sequence. Research suggest that establishing classroom norms for productive discourse is essential (Mercer et al., 2004; Michaels et al., 2008) but beyond that little knowledge exists that could reliably guide action. One solution is an attempt to identify "high-leverage practices" (Windschitl, 2009) but little evidence exists as to what those might be. Waiting more than a second for a student answer to a teacher question is one well-known practice that has been identified to have a significant positive effect on student learning. Yet knowledge of its value is far from pervasive and its use even less common. The medical equivalent is the regular and thorough washing of hands. Yet 150 years after Lister's identification of its importance, it still proves difficult to implement (Gawande, 2007).

More importantly for teacher learning, is the need to recognize that failure is an inherent feature of acquiring the new skill. Teachers on professional development programs must, therefore, be prepared for failure and how to deal with its consequences both for themselves and their students. Progress can only be attained by reflection on practice and identifying the flaw or flaws that led to failure. Skilled coaches who can help identify the current weakness in practice help enormously at this point. Likewise, recognizing aspects that contribute to success with any practice that needs to be consolidated is equally important.

These reflections lead to a view that the essential elements of any professional development which seeks to change teacher practice are:

- A clear and well-articulated rationale for the practice with supporting empirical evidence of its value.
- Video exemplars of expert use of the practice.
- Identification of key features of the practice and a sequence of progressive steps towards its attainment.
- A mechanism for providing feedback helping teachers to identify both success and failure.
- Extensive opportunities to practice.

## Limitations of the Study

There are a number of limits to this study that the reader needs to be cognizant of interpreting the findings and their implications. First is that we have no exact measure of the nature of the intervention. We do not know which students were taught by which teachers, how many lessons were devoted to the use of argumentation, how long these experiences were and their nature. What we do have for each school is a map of how the curriculum was affected, the manner in which the intervention was implemented, the nature and focus of the discussions at the reflective meetings and a set of video observations of the lead teachers and others teachers who agreed to be videoed. However, these videos cannot be considered as representative of what was, or was not happening within the department as a whole. Taken together all these data have enabled us to build a picture of the differing ways in which the intervention was enacted in the four schools. Full details will be reported in future publications.

The consequence is that the intervention must be considered to be a treatment which is happening at the level of the school and the study one which has attempted to measure its outcomes on students as whole rather than a model which sought to measure the effects at the teacher level and then on students—a design which would have enabled analysis using hierarchical linear modeling. A design, however, that would have required considerable more resource to undertake.

Second is that we have no direct measure of students' facility with argumentation in science—the capability that the intervention was trying to develop. Specific tests that do exist of scientific reasoning such as the Lawson test (Lawson, 1978) have been designed to test students facility with logico-mathematical reasoning and not their ability to construct and critique evidence-based arguments reliant on data and justifiable warrants. The lack of such instruments remains a serious problem for quantitative intervention studies which focus on argumentation. Whether the construct measured by Raven's matrices is a justifiable proxy for students' ability with argumentation is uncertain.

Thus, what this study provides is essentially a "fuzzy generalization"—a summary which states that "particular events may lead to particular consequences," or in this case to no consequences, and which depends on making "clear the context of the statement and the evidence justifying it," in so doing it provides "a powerful and user-friendly summary which can serve as a guide to professional action" (Bassey, 2001). What lends weight to our conclusions is that it is drawn from a number of measures of different constructs and in no case was there any evidence of a significant effect. It is this absence of an effect that warrants publication and exploration of its implications for the field.

## Conclusions and Implications

This research contributes to informing the work on professional development in two ways. Two possible hypotheses could explain the findings. First, the minimal level of intervention used in this research may be insufficient to achieve any substantive change in teacher knowledge and practices—that is change may only be achieved by more sustained and intensive professional development that requires more resource either of teacher time, peer support, or specific coaching. Second, it raises the question of whether effects might have been seen in the years after the intervention as has been observed in other studies when teachers have fully assimilated the practice. Eliminating this second hypothesis as an explanation in future studies will require extended research designs that look for student effects in the years following the intervention.

Third, what these findings do is add to the body of literature that suggests that teacher change is a complex and slow process. As Opfer and Pedder (2011) point out the relationship between changes in practice, changes in belief, and changes in students is reciprocal—with changes in one being contingent on changes in another. As a corollary, teacher learning is essentially complex and it requires a more holistic approach to its study that should aim at "identifying the edges of generalizability and variation that characterize the patterns and process that constitute teacher learning (p. 396)." What this study offers then is a contribution to identifying one such edge—that is the minimal level of intervention, style of leadership and time that is necessary to see the kind of change in practice that can have a measurable impact on student outcomes. Moreover, we would go further in arguing that just as the lessons learned from the disasters of engineering design have done more to advance engineering knowledge than all the successes (Johnson, 2009; Petroski, 1996), that likewise there is much to be learnt from those attempts at professional development which do not succeed. Identifying failure requires clear and valid measures of outcomes that must assess at a minimum changes in teacher performance. Moreover, it also requires a clear and theoretically well-grounded model underlying its design which can be examined for the flaws and mistaken assumptions that led to less than optimal outcomes. In presenting the outcomes of this work, we have attempted to offer a model for testing and evaluating professional development from which the community as a whole might learn. In that sense, the research is a contribution to identifying the lessons of the past so that others are not condemned to repeat them.

If we can eliminate our second hypothesis that no effect was seen because the intervention had insufficient time to take root and for teachers to build their competence with argumentation, then the lack of any student effects would suggests that transforming teacher practice is not amenable to such minimal solutions. Rather, it suggests that change will only be achieved with more sustained and intensive professional development that requires more resource of teacher time, peer support, or coaching and models of what are the major identifiable steps in pedagogic practice with argumentation in school science. Essentially, the field needs a deeper understanding of the nature of professional development and instructional strategies that will lead to greater use by teachers of science of a more dialogic approach. In particular, what are the critical features of any program and the key processes that will initiate and mediate change? Only when we have systematic and rigorous answers to these questions will it be worth exploring the effects on student outcomes.

Commenting on the nature of research evidence and the implications of negative findings, Millar and Osborne (2009) make the following point:

> What if the outcome of the trial is less positive, that is, there is no significant difference in average score between the experimental and control groups? Can we conclude that the intervention is not worth using? ... Perhaps the underlying strategy on which the experimental intervention is based is sound, but it has not been implemented well with this particular content. Or perhaps the lack of commitment to the intervention by some of those required to implement it has cancelled out any gains achieved by those who did favor it.

Whilst the data collected from this study cannot provide definitive answers to such questions, they do go some way to filling some of the vacuum of empirical evidence about interventions at scale, using participants who were not all volunteers that have attempted to measure the effect on student learning, reasoning, and engagement. As Roth et al. (2011) point out, "surprisingly little has been documented about the impact of professional development on science teacher learning, teaching practice, and student learning" (p. 117) and then most of that has been on teacher learning or teacher practice rather than student learning. Only recently has work begun to appear which addresses this lacuna. This work is one contribution to our knowledge of what needs to be done to effect teacher practice and, in turn, student learning. While our project may have been a bridge too far we hope that it offers valuable insights for building and deepening our understanding of what can and cannot be achieved by teacher professional development.

This work was funded by the UK Economic and Social Science Research Council - Grant No RES-062-23-0571.

## Notes

[1] As a matter of record and important to interpreting some of the positions adopted in this article, the work reported here was originally conceived and conceptualized in 2006.

[2] Author's original emphasis.

[3] The UK education system refers to Years rather than Grades. Children in Year 7 are commonly 11–12 years old and in Year 10, 15–16 years old.

[4] We use the term here ''authentic'; in the sense used by Sartre (1969) in that it is not a property of the context but an outcome of a judgment by each individual about the setting and the actions they observe.

## References

Adey, P., & Shayer, M. (1993). An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. Cognition & Instruction, 11(1), 1.

Alexander, R. (2005). Towards dialogic teaching. York: Dialogos.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. Science, 333(6045), 1034–1037.

Alverman, D. E., Qian, G., & Hynd, C. E. (1995). Effects of interactive discussion and text type on learning counterintuitive science concepts. Journal of Educational Research, 88, 146–154.

Ames, G. J., & Murray, F. B. (1982). When two wrongs make a right: Promoting cognitive change by social conflict. Developmental Psychology, 18, 894–897.

Andrews, R. (1995). Teaching and learning argument. London: Cassell.

Applebee, A., Langer, J., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. American Educational Research Journal, 40(3), 685.

Avvisati, F., & Vincent-Lancrin, S. P. (in press). Effective teaching for improving students' motivation, curiosity, and self-confidence in science: A comparative approach. Paris: Centre for Educational Research and Innovation, OECD.

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.), Multiple perspectives on the teaching and learning of mathematics (pp. 83–104). Wesport, CT: Ablex.

Bassey, M. (2001). A solution to the problem of generalisation in educational research: Fuzzy prediction. Oxford Review of Education, 27(1), 5–22.

Billig, M. (1996). Arguing and thinking (2nd ed.). Cambridge: Cambridge University Press.

Birman, B. F., Le Floch, K. C., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., ... Yoon, K. S. (2007). *State and Local Implementation of the "No Child Left Behind Act."* Volume II—Teacher Quality under"NCLB": Interim Report.

Blalock, C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. (2008). In pursuit of validity: A comprehensive review of science attitude instruments. International Journal of Science Education, 30(7), 961–977.

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: The case of Railside School. The Teachers College Record, 110(3), 608–645.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. Educational Researcher, 33(8), 3–15.

Brown, A. L., & Campione, J. C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), Classroom lessons: Integrating cognitive theory and classroom practice (pp. 229–270).

Cerini, B., Murray, I., & Reiss, M. (2003). Student review of the science curriculum. London: NESTA.

Chi, M. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science, 1, 73–105.

Christodoulou, A. (2011). *The science classroom as a site of epistemic talk: two case studies of teachers and their students*. Unpublished PhD Thesis. King's College London.

Clark, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. Teaching and Teacher Education, 18, 947–967.

Claxton, G. (1988). Live and Learn: An introduction to the psychology of growth and change in everyday life. Milton Keynes: Open University Press.

Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., et al. (1991). Assessment of a problem-centered second-grade mathematics project. Journal for Research in Mathematics Education, 22, 3–29.

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. Contemporary Educational Psychology, 29(2), 186–204.

Csikszentmihalyi, M., & Hermanson, K. (1995). Intrinsic motivation in museums: Why does one want to learn? In J. H. Falk & L. D. Dierking (Eds.), Public institutions for personal learning: Establishing a research agenda (pp. 67–77). Washington, DC: American Association of Museums.

Csikszentmihalyi, M., & Schneider, B. (2000). Becoming adult: Preparing teenagers for the world of work. New York: Basic Books.

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

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. Educational Researcher, 38(3), 181–199.

Dillon, J., Osborne, J. F., Fairbrother, R., & Kurina, L. (2000). A study into the professional views and needs of science teachers in primary & secondary schools in England. London: King's College London.

Doise, W., & Mugny, G. (1984). The social development of the intellect. Oxford: Pergamon.

Driver, R., Leach, J., Millar, R., & Scott, P. (1996). Young people's images of science. Buckingham: Open University Press.

Driver, R., Newton, P., & Osborne, J. F. (2000). Establishing the norms of scientific argumentation in classrooms. Science Education, 84(3), 287–312.

Dweck, C. (2000). Self-theories: Their role in motivation, personality, and development (essays in social psychology). Philadelphia, PA: Psychology Press.

Elder, A. D. (2002). Characterizing fifth grade students' epistemological beliefs in science. In B. K. Hofer & P. R. Pintrich (Eds.), Personal epistemology (pp. 347–364). Mahwah, New Jersey: Lawrence Erlbaum.

Ennis, R. H., Millman, J., & Tomko, T. (1985). Cornell critial thinking tests, level X and Z. Pacific Grove, California: MidWest Publications.

Erduran, S., & Jiménex-Aleixandre, M. P. (2008). Argumentation in science education: Perspectives from classroom-based research. Dordrecht: Springer.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100(3), 363.

Ford, M. J. (2008). Disciplinary authority and accountability in scientific practice and learning. Science Education, 92(3), 404–423.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., et al. (2011). Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation (NCEE 2011-4024). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of. Education Sciences, U.S.

Gawande, A. (2007). Better: A surgeon's notes on performance. New York: Henry Holt.

Goldberg, T., Schwarz, B. B., & Porat, D. (2008). Living and dormant collective memories as contexts of history learning. Learning and Instruction, 18, 223–237.

Graff, G., & Birkenstein, C. (2007). They say/I say: The moves that matter in persuasive writing: WW Norton & Company Incorporated.

Greenleaf, C., Litman, C., Hanson, T. L., Rosen, R., Boscardin, C. K., Herman, J., et al. (2011). Integrating literacy and science in biology: Teaching and learning impacts of reading apprenticeship professional development. American Educational Research Journal, 48(3), 647–717.

Greeno, J. (1998). The situativity of knowing, learning and research. American Psychologist, 53(1), 5–26.

Guskey, T. R., & Yoon, K. S. (2009). What works in professional development. Phi Delta Kappan, 90(7), 495–500.

Hannover, B., & Kessels, U. (2004). Self-to-prototype matching as a strategy for making academic choices. Why high school students do not like math and science. Learning and Instruction, 14(1), 51–67.

Harré, R. (1984). The philosophies of science: An introductory survey (2nd ed.). Oxford: Oxford University Press.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. Journal of Research in Science Teaching, 49(3), 333–362, DOI: 10.1002/tea.21004

Hoban, G. (2002). Teacher Learning for Educational Change. Buckingham: Open University Press.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. Review of Educational Research, 67(1), 88–140.

Horton, R. (1967). African traditional thought and western science. Africa, 37, 157.

Howe, C., Tolmie, A., & Mackenzie, M. (1995). Computer support for collaborative learning of physics concepts. In C. O'Malley (Ed.), Computer-supported collaborative learning. Berlin: Springer.

Howson, C., & Urbach, P. (2006). Scientific reasoning: A Bayesian approach (3rd ed.). Chicago: Open Court.

Hynd, C., & Alvermann, D. E. (1986). The role of refutation text in overcoming difficulty with science concepts. Journal of Reading, 29(5), 440–446.

Johnson, A. (2009). Hitting the brakes: Engineering design and the production of knowledge. Durham: Duke University Press.

Khine M. (Ed.). (2011). Perspectives on scientific argumentation: Theory, practice and research. Dordrecht: Springer.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. Science, 333, 971–975.

Kuhn, D. (1992). Thinking as argument. Harvard Educational Review, 62(2), 155–178.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. Science Education, 77(3), 319–337.

Kuhn, D. (1999). A developmental model of critical thinking. Educational Researcher, 28(2), 16–46.

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. Cognitive Development, 15, 309–328.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. American Educational Research Journal, 27(1), 29–63.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. Journal of Research in Science Teaching, 15(1), 11–24.

Lederman, N. G., & O'Malley, M. (1990). Students' perceptions of tentativeness in science: Development, use, and sources of change. Science Education, 74, 225–239.

Lee, M.-H., Wu, Y.-T., & Tsai, C.-C. (2009). Research Trends in Science Education from 2003 to 2007: A content analysis of publications in selected journals. International Journal of Science Education, 31(15), 1999–2020.

Lehrer, J. (2010). *The truth wears off.* New Yorker, 52–57.

Lord Sainsbury of Turville. (2007). The race to the top: A review of government's science and innovation policies. London: HM Treasury.

Loucks-Horsley, S., Hewson, P., Love, N., & Stiles, K. E. (2003). Designing professional development for teachers of science and mathematics (2nd ed.). Thousand Oaks, California: Corwin Press, Inc.

Martin, A., & Hand, B. (2009). Factors affecting the implementation of argument in the elementary science classroom. A longitudinal case study. Research in Science Education, 39(1), 17–38.

McLaughlin, M. W., & Talbert, J. E. (1993). Teaching for understanding: Challenges for policy and practice. San Francisco, CA: Jossey-Bass.

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. British Education Research Journal, 30(3), 359–377.

Mercer, N., Wegerif, R., & Dawes, L. (1999). Children's talk in the development of reasoning. British Educational Research Journal, 25(1), 95–111.

Michaels, S., O'Connor, C., & Resnick, L. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. Studies in the Philosophy of Education, 27, 283–297.

Millar, R., & Osborne, J. F. (2009). Research and practice: A complex relationship? In M. C. Shelley II, L. Yore, & B. Hand (Eds.), Quality research in literacy and science education. (pp. 41–61). Dordrecht: Springer.

Moller, A. P., & Jennions, M. D. (2001). Testing and adjusting for publication bias. Trends in Ecology & Evolution, 16(10), 580–586.

National Academies of Sciences. (2010). Rising above the gathering storm revisited. Washington, DC: National Academy of Sciences.

National Research Council. (2007). Taking science to school: Learning and teaching in grades K-8. Washington, DC: National Research Council.

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: Committee on a Conceptual Framework for New K-12 Science Education. Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education.

Newton, P., Driver, R., & Osborne, J. F. (1999). The place of argumentation in the pedagogy of school science. International Journal of Science Education, 21(5), 553–576.

Nolen, S. B. (2003). Learning environment, motivation and achievement in high school science. Journal of Research in Science Teaching, 40(4), 347–368.

Nystrand, M., Gamoran, A., Kachur, R., & Prendegarst, C. (1997). Opening dialogue: Understanding the dynamics of language and learning in the English classroom. New York: Teachers College Press.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. Review of Educational Research, 78(1), 33–84.

Oaksford, M., & Chater, N. (2007). Bayesian Rationality: The probalistic approach to human reasoning. New York: Oxford University Press.

Ogborn, J. (2002). Ownership and transformation: Teachers using Curriculum Innovation. Physics Education, 37(2), 142–146.

Opfer, V. D., & Pedder, D. (2011). Conceptualizing teacher professional learning. Review of Educational Research, 81(3), 376–407.

Osborne, J. F., & Collins, S. (2001). Pupils' views of the role and value of the science curriculum: A focus-group study. International Journal of Science Education, 23(5), 441–468.

Osborne, J. F., Erduran, S., & Simon, S. (2004a). Enhancing the quality of argument in school science. Journal of Research in Science Teaching, 41(10), 994–1020.

Osborne, J. F., Erduran, S., & Simon, S. (2004b). The IDEAS project. London: King's College London.

Paris, S. G. (1998). Situated motivation and informal learning. Journal of Museum Education, 22(2 & 3), 22–26.

Pell, T., & Jarvis, T. (2001). Developing attitude to science scales for use with children of ages from five to eleven years. International Journal of Science Education, 23(8), 847–862.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? Educational Researcher, 18(1), 16–25.

Perry, W. G. (1970). Forms of intellectual and ethical development in the college years. A scheme. New York: Holt, Rinehart and Winston.

Petroski, H. (1996). Engineering by design: How engineers get from thought to thing. Cambridge: Harvard University Press.

Pontecorvo, C. (1987). Discussing and reasoning: The role of argument in knowledge construction. In E. De Corte, H. Lodewïjks, R. Parmentier, & P. Span (Eds.), Learning and instruction: European Research in an International Context (pp. 239–250). Oxford: Pergamon Press.

Pontecorvo, C., & Girardet, H. (1993). Arguing and Reasoning in understanding historical topics. Cognition and Instruction, 11(3 & 4), 365–395.

Ratcliffe, M., Hanley, P., & Osborne, J. F. (2007). Evaluation of Twenty First Century Science GCSE Strand 3: The teaching of Twenty First Century Science GCSE, and teachers' and students' views of the course. Southampton: University of Southampton.

Raven, J., Raven, J. C., & Court, J. H. (updated 2004). (2004). Manual for Raven's Progressive Matrices and Vocabulary Scales. San Antonio, TX: Harcourt Assessment.

Renninger, K. A., Hidi, S., & Krapp, A. (1992). The role of interest in learning and development. Hillsdale, NJ: Lawrence Erlbaum.

Resnick, L., Michaels, S., & O'Connor, C. (in press). How (well-structured) talk builds the mind. In R. Sternberg & D. Preiss (Eds.), From genes to context: New discoveries about learning from educational research and their applications. New York: Springer.

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Videobased lesson analysis: Effective science PD for teacher and student learning. Journal of Research in Science Teaching, 48(2), 117–148, DOI: 10.1002/tea.20408

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. Journal of the Learning Sciences, 12(1), 5–51.

Sandoval, W. A., & Morrison, K. (2003). High school students' ideas about theories and theory change after a biological inquiry unit. Journal of Research in Science Teaching, 40(4), 369–392.

Sandoval, W. A., & Millwood, K. (2005). The quality of students' use of evidence in written scientific explanations. Cognition and Instruction, 23(1), 23–55.

Sartre, J.-P. (1969). Being and Nothingness. Northampton: John Dickens and Co.

Sfard, A. (2008). Thinking as communicating: Human development, the growth of discourses, and mathematizing. Cambridge, UK: Cambridge University Press.

Shayer, M., Wylam, H., Adey, P., & Kuchemann, D. (1979). *CSMS Science Reasoning Tasks*. NFER, Nelson, Windsor.

Simon, S., Erduran, S., & Osborne, J. F. (2006). Learning to teach argumentation: Research and development in the science classroom. International Journal of Science Education, 28(2–3), 235–260.

Sisk-Hilton, S. (2009). Teaching and learning in public. New York: Teachers College Press.

Spillane, J. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. Journal of Curriculum Studies, 31(2), 143–175.

Spillane, J. (2006). Distributed leadership. San Francisco: John Wiley & Sons.

Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2008). Orchestrating productive mathematical discussions: Five practices for helping teachers move beyond show and tell. Mathematical Thinking and Learning, 10(4), 313–340.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. Journal of the American Statistical Association, 54, 30–34.

Szu, E., & Osborne, J. F. (2011). Scientific reasoning and argumentation from a Bayesian perspective. In M. S. Khine (Ed.), Perspectives on scientific argumentation (pp. 55–71). Dordrecht: Springer.

Taconis, R., & Kessels, U. (2009). How choosing science depends on students' individual fit to 'Science Culture'. International Journal of Science Education, 31(8), 1115–1132.

Tytler, R., Osborne, J. F., Williams, G., Tytler, K., Clark, J. C., Tomei, A., et al. (2008). Opening up pathways: Engagement in STEM across the Primary-Secondary school transition. A review of the literature concerning supports and barriers to Science, Technology, Engineering and Mathematics engagement at Primary-Secondary transition. Commissioned by the Australian Department of Education, Employment and Workplace Relations. Melbourne: Deakin University.

van Lier, L. (1996). Interaction in the language curriculum. New York: Longman.

Venville, G. J., & Dawson, V. M. (2010). The impact of a classroom intervention on grade 10 students' argumentation skills, informal reasoning, and conceptual understanding of science. Journal of Research in Science Teaching, 47(8), 952–977.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? Reading in a Foreign Language, 15(3), 130–163.

Watson, J. D., & Crick, F. H. C. (1953). A Structure for deoxyribose nucleic acid. Nature, 171, 737–738.

Weiss, I. R., Pasley, J. D., Sean Smith, P., Banilower, E. R., & Heck, D. J. (2003). A Study of K–12 Mathematics and Science Education in the United States. Chapel Hill, NC: Horizon Research.

Wertsch, J. (1991). Voices of the Mind: A sociocultural approach to mediated action. Cambridge, MA: Harvard University Press.

Windschitl, M. (2009). *Cultivating 21st Century Skills in Science Learners: How Systems of Teacher Preparation and Professional Development will have to evolve*. Paper presented at the National Academies of Science Workshop on 21st Century Skills.

Yoon, K. S., Ducan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement. Washington, DC: US Department of Education, Institute of Education Sciences.

Zohar, A. (2004). Higher order thinking in science classrooms: Students' Learning and Teachers Professional Development. Dordrecht: Kluwer.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. Journal of Research in Science Teaching, 39(1), 35–62.