Education
Endowment
Foundation

Philosophy for Children

Evaluation report and Executive summary

July 2015

Independent evaluators:

Stephen Gorard, Nadia Siddiqui and Beng Huat See
(Durham University)

Durham
University

The Education Endowment Foundation (EEF)

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- Identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- Evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale;
- Encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust, as lead charity in partnership with Impetus Trust (now part of Impetus-The Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:

Robbie Coleman
Research and Communications Manager
Education Endowment Foundation
9th Floor, Millbank Tower
21-24 Millbank
SW1P 4QP

p: 020 7802 1679
e: robbie.coleman@eefoundation.org.uk
w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from Durham University led by Professor Stephen Gorard.

Stephen Gorard is Professor of Education and Well-being, and Fellow of the Wolfson Research Institute at Durham University. He is a Methods Expert for the US government Institute of Education Science, member of the ESRC Grants Awarding Panel, and Academician of the Academy of Social Sciences.

His work concerns the robust evaluation of education as a lifelong process, focused on issues of equity and effectiveness. He regularly advises governments and other policy-makers, including oral and written advice to the House of Commons Education Committee every year since 2003. He is also a widely read and cited methodologist, involved in international and regional capacity-building activities, and used regularly as an adviser on the design of evaluations by central and local governments, NGOs and charities. He is currently an evaluator for the European Commission Directorate-General for Education and Culture, the Department of Work and Pensions, the Food Standards Agency, the Learning and Skills Information Service, and the Education Endowment Foundation. He is author of nearly 1,000 books and papers.

Contact details:

Professor Stephen Gorard
School of Education
Durham University
Leazes Road
DH1 1TA

p: 0191 334 8419
e: s.a.c.gorard@durham.ac.uk

Contents

Executive summary

The project

Philosophy for Children (P4C) is an approach to teaching in which students participate in group dialogues focused on philosophical issues. Dialogues are prompted by a stimulus (for example, a story or a video) and are based around a concept such as 'truth', 'fairness' or 'bullying'. The aim of P4C is to help children become more willing and able to ask questions, construct arguments, and engage in reasoned discussion.

The primary goal of this evaluation was to assess whether a year of P4C instruction for pupils in Years 4 and 5 would lead to higher academic attainment in terms of maths, reading, and writing. The project also assessed whether P4C instruction had an impact on Cognitive Abilities Test results.

The evaluation ran from January to December 2013. Teachers were trained in P4C by the Society for the Advancement of Philosophical Enquiry and Reflection in Education (SAPERE). On average, pupils received one period of P4C per week, although this varied across schools. A total of 48 schools across a wide range of English geographies participated. While these schools were in many ways diverse, as a whole they had above-average levels of disadvantaged pupils.

The project was delivered by SAPERE, funded by the Education Endowment Foundation, and independently evaluated by a team at Durham University.

Key Conclusions
1. There is evidence that P4C had a positive impact on Key Stage 2 attainment. Overall, pupils using the approach made approximately two additional months' progress in reading and maths.
2. Results suggest that P4C had the biggest positive impact on Key Stage 2 results among disadvantaged pupils (those eligible for free school meals).
3. Analyses of the Cognitive Abilities Test (a different outcome measure not explicitly focused on attainment) found a smaller positive impact. Moreover, in terms of this outcome it appears that disadvantaged students reaped fewer benefits from P4C than other pupils. It is unclear from the evaluation why there are these differences between the two outcomes.
4. Teachers reported that the overall success of the intervention depended on incorporating P4C into the timetable on a regular basis. Otherwise there was a risk that the programme would be crowded out.
5. Teachers and pupils generally reported that P4C had a positive influence on the wider outcomes such as pupils' confidence to speak, listening skills, and self-esteem. These and other broader outcomes are the focus of a separate evaluation by the University of Durham.

Security rating

Security rating awarded as part of the EEF peer review process

Findings have a moderate degree of security. The evaluation was set up as a randomised controlled trial with school-level randomisation. The study was classified as an 'effectiveness trial', meaning that it sought to test whether the intervention can work at scale. Before the trial started, there was a good balance of pupil characteristics between the group receiving P4C and those in the comparison group. No schools dropped out of the trial, and all results for participating students were available for the primary outcome (Key Stage 2 results).

Results

- The evaluation found evidence that P4C had a positive impact on pupils' Key Stage 2 (KS2) progress in reading and maths. This is significant in that P4C was not explicitly focused on improving KS2 outcomes, yet managed to lift maths and reading attainment relative to 'business as usual'.

- It is also important to note that the gains in KS2 were greater in all subjects for students eligible for free school meals (FSM).
- Results on the Cognitive Abilities Test (CAT) showed mixed results. Pupils who started the programme in Year 5 showed a positive impact, but those who started in Year 4 showed no evidence of benefit. Taken together, the results suggested that P4C resulted in a small improvement in CAT scores; however, this should be treated with caution. FSM-eligible pupils did not appear to benefit relative to a comparison group of FSM-eligible pupils who received normal lessons rather than P4C.
- All of the gain in the CAT scores comes from the verbal subscale. There was very little difference between treatment and control groups in terms of quantitative, non-verbal, and spatial elements of the CAT.
- The duration of the programme – which exposed pupils to P4C for just under a year – may not have been long enough for the full impact of P4C to be felt. Previous UK research in which larger effect sizes were found examined a 16-month period (see Topping and Trickey 2007).
- There was no evidence of improvement at Key Stage 2 for writing. This was not a surprise, as the programme did not involve writing skills. It is worth noting that the writing results of the P4C pupils improved at least as much as those of pupils who received normal classes.
- Teachers reported that the overall success of the intervention depended on incorporating P4C into the timetable on a regular basis, or there was a risk that the programme would be crowded out by activities that appear to more directly address the national curriculum.
- Feedback from teachers and pupils also suggested a belief that P4C had a beneficial impact on wider outcomes such as pupils' confidence to speak, patience when listening to others, and self-esteem. Some teachers also perceived that P4C had a positive impact on general classroom engagement and may have resulted in some pupils asking more questions across all lessons.
- These and other broader outcomes are the focus of a separate evaluation by the University of Durham (see http://www.nuffieldfoundation.org/non-cognitive-impacts-philosophy-children).

Cost

Financial costs for the programme were largely for teacher training. This involved two days of training before P4C was delivered, and ongoing support throughout the year. There were also small costs for stimulus books, website subscription, and SAPERE membership. Overall, the total financial cost to deliver this project in one school for one academic year was £3,940, or roughly £16 per pupil per year for a school of 240 pupils. Advanced level teacher's training would cost more, and is now recommended by SAPERE for at least some teachers in each school. However, the impact of this level of training was neither tested nor costed in this evaluation. The programme imposed some costs in terms of teacher time. Two days were required for the initial SAPERE training. Teachers reported that additional time was necessary for them to prepare P4C lessons, although it was not clear whether this was simply the extra time involved because P4C was new to them.

Test domain	Key Stage 2 test domain	Effect size	Months' Progress	Security Rating	Cost
P4C vs comparison group	Maths	+0.10	+2 months		£
	Reading	+0.12	+2 months		
	Writing	+0.03	0 months		
Free School Meal Pupils: P4C vs comparison group	Maths	+0.20	+3 months	NA	
	Reading	+0.29	+4 months		
	Writing	+0.17	+2 months		

Introduction

Intervention

The intervention is known as 'Philosophy for Children' (P4C) – an educational approach centred on nurturing philosophical enquiry. The aim of the programme is to help children become more willing and able to question, reason, construct arguments and collaborate with others. Through the training and development of teachers, the initiative is intended to foster cognitive improvement and greater self-confidence in young people, leading to higher academic attainment and non-cognitive development in areas such as pupils' self-esteem.

Background

Philosophy for Children (P4C) was originally developed by Professor Matthew Lipman in New Jersey, USA in 1970 with the establishment of the Institute for the Advancement of Philosophy for Children (IAPC). This organisation popularised and developed the idea of teaching thinking skills at school level through the medium of philosophical dialogue. In 1990 the BBC televised a documentary, 'Socrates for 6 year olds' which showed P4C being practised in one of the most challenging schools in Merrick, USA (Lipman 1990). This documentary generated a lot of interest among UK academics, school teachers, education service providers, and charities, which culminated in the establishment of the Society for the Advancement of Philosophical Enquiry and Reflection in Education (SAPERE) in 1992 (http://p4c.com/history-p4c). SAPERE, a non-profit society, promotes the use of P4C in UK schools along with developing teaching resources and providing teacher training courses. P4C is practised across all education age ranges.

SAPERE's model of P4C differs in some ways from Lipman's original conception. In particular, there is no use of specially written philosophical novels. Materials recommended by SAPERE include stories, poems, scripts, short films, images, artefacts, and picture books. However, Lipman's central aim of creating a classroom 'community of enquiry' is retained along with the broad sequence of activities that constitute a P4C session.

P4C has become a worldwide approach and has been adopted by schools in 60 countries across the world, although the nature of the practice varies. It is difficult to estimate the number of schools in the UK using P4C. However, SAPERE as a leading service provider claims to have 600 registered schools that have been regularly implementing P4C.

Lipman and his colleagues carried out an initial evaluation of P4C (Lipman et al. 1980). This was a small study using a pre- and post-test experimental design involving a total of 40 pupils from two schools in the Montclair District of New Jersey. The research design matched 20 pupils who received the intervention with a control group of 20 pupils, though it is not clear how matching was done. The intervention group was taught P4C twice a week for nine weeks by Lipman, while the control group was taught social studies. The study reported significant gains in logical reasoning and reading, measured using the California Test of Mental Maturity (CTMM). Significant differences in reading scores were reported to be maintained 2.5 years later. A second, larger experiment reported in Trickey and Topping (2004), involved 200 pupils (100 in each arm). This time lessons were taken by teachers over a period of two years. The authors reported significant improvements in reading and critical thinking, but the outcomes for logical thinking and the use of questions were unclear. It was not clear how schools were selected and groups matched, neither was there any report of attrition (Trickey and Topping 2004).

Six other studies reviewed by Trickey and Topping (2004) showed consistent moderate effects on a range of outcome measures. The mean effect size for the review was 0.43. However, these studies were not always comparable because of the different outcomes measured and the different

instruments used for measuring them. For example, Fields (1995) and IAPC (2002) used the New Jersey Test of Reasoning Skills (NJTRS), while Doherr (2000) assessed emotional intelligence using a Cognitive Behavioural Therapy Assessment. Campbell (2002) evaluated listening and talking skills using questionnaires, focus groups, interviews, and observations. It has to be noted that the NJTRS was specially developed for Lipman and the IAPC to measure reasoning skills taught in the P4C curriculum. This is likely to bias the results against the control group of pupils not exposed to the P4C curriculum. Moriyon and Tudela (2004) noted that studies using NJTRS showed larger effect sizes than tests of literacy and numeracy.

One of the earliest studies in the UK was conducted by Williams (1993) (who is also the programme developer for the current evaluation funded by the Educational Endowment Foundation). The 1993 study examined the effects of 27 one-hour P4C lessons (using Lipman's materials) on reading comprehension, reasoning skills, and intellectual confidence. A total of 42 pupils from two Year 7 classes in one school in Derbyshire took part, with results reported for 32 children. Participants were randomised to receive P4C lessons (n = 15) or extra English (n = 17). Pre- and post-test comparison of reading comprehension using the London Reading Test showed that P4C pupils made significantly bigger gains than control pupils. Significant gains were also reported for reasoning skills and intellectual confidence. These were measured using both bespoke evaluation tools and video recordings of pupils' interaction during lessons; the evaluators made subjective judgements on the latter. The study showed that the philosophy group registered improvements in reasoning behaviour, while the control group showed no such improvements.

Mercer et al. (1999) evaluated the impact of the Talk, Reasoning, & Computers (TRAC) programme which trained pupils to follow certain ground rules for collaborative talking of the kind necessary to implement P4C in a primary classroom. It consisted of nine structured teacher-led lessons of collaborative activities, including some that were computer-based, carried out over ten weeks. The study involved 60 Year 4 and 5 pupils (ages 9 and 10) from three middle schools in Milton Keynes, UK. Each lesson was one hour long. Pupils' reasoning abilities were assessed using the Raven's Progressive Matrices test of non-verbal reasoning. Observational data and pupils' interactions were also recorded. Pupils receiving the programme made significantly bigger gains between pre- and post-test compared to control pupils.

Despite this evidence of short-term improvements, some commentators suggested that the effects of the programme may not be immediately obvious because of the difficulty of finding a valid and reliable instrument sensitive enough to measure short-term changes in reasoning skills (Adey and Shayer 1994).

The longer-term impact of P4C was assessed by Topping and Trickey (2007). They followed pupils over two years. A total of 177 pupils (105 experimental and 72 control) from eight schools and eight classes in Dundee, UK were matched and randomised. Pupils were tracked from the penultimate year of primary school to the first year of secondary school. Pupils' cognitive abilities were measured using the CAT. Complete data was available for only 115 pupils. Experimental pupils received 1 hour per week of collaborative enquiry lesson, while control pupils continued regular lessons. After 16 months of intervention (with 1 hour of P4C per week) the treatment pupils had made substantial improvements in test scores whereas control pupils performed worse on post-test than on pre-test (ES = 0.7). Results two years later indicated that treatment pupils maintained their advantage in follow-up test scores compared to the control pupils. The intervention effect for the CAT score appeared to be maintained for the more able pupils in the follow-up, but not for the lowest-achieving pupils (Topping and Trickey 2007, table 3, p. 794).

Some commentators have suggested that the effects of the programme may not be immediately obvious because of the difficulty in finding a valid and reliable instrument sensitive enough to measure short-term changes in reasoning skills (Adey and Shayer 1994).

A more recent longitudinal study of the long-term impact of P4C was conducted in Madrid (Colom et al. 2014). This was intended to track children attending two private schools over 20 years. A total of 455 children aged 6 (first year of primary school) to 18 (final year of high school) from one school were trained in the P4C programme. Another 321 pupils from another school matched on demographic characteristics formed the control group. Data on children's cognitive, non-cognitive, and academic achievements were collected at three time points, at ages 8, 11/12, and 16. Preliminary analyses of 281 treatment children and 146 control children showed that the programme had positive impacts on general cognitive ability (ES = 0.44), but results on academic achievement were not yet available. The authors implied that the programme was particularly beneficial for lower-ability pupils, but this was not clear from their presentation of the analysis. Moreover, although the study was large scale and long term, pupils were not randomised in terms of receiving P4C instruction, and the study may not be generalisable as participants came from relatively prosperous families. In short, the results from this preliminary analysis should be treated with a high degree of caution.

Many of the studies so far have used a matched comparison design, and most have measured cognitive abilities, reasoning skills or other affective outcomes rather than school attainment directly. Moreover, while there have been several studies in the UK, they have tended to be small scale. It is therefore difficult to say whether philosophical enquiry can lead to enhanced performance in academic domains and whether it would have the same impact on British children in UK schools. No proper large-scale randomised controlled trial has been conducted on this as far as we know. There are some unsystematic observations of beneficial impact from OFSTED reports. There is therefore evidence of promise in terms of raising attainment but this has not yet been tested robustly at scale, and so a large trial is appropriate.

Objectives

The main aim of the impact evaluation was to determine the effect of the P4C programme on the Key Stage 2 scores of pupils who were in Year 5 when the schools were randomised and Year 6 by the end of the trial. A further outcome was the scores on the CAT4 (Cognitive Ability Test, Fourth Edition; GL Assessment website) for those in Years 4 and 5 at the start and in Years 5 and 6 when post-tested. A subsequent aim proposed by the developers and agreed by the funders was to consider the short-term impact of P4C on pupils' experiences at school, and their relationship with teachers and peers.

The process evaluation was designed to assess fidelity to treatment and to collect the views of teachers, school staff members, and pupils regarding P4C impact and implementation.

Project team

The programme was developed by Society for the Advancement of Philosophical Enquiry and Reflection in Education (SAPERE); the society delivered the training to teachers and provided ongoing support to teachers and schools.

In each participating school a P4C lead staff member was appointed through agreement with the SAPERE team. The lead was responsible for the implementation and communication with the SAPERE team and evaluators. The P4C school leads were usually the literacy coordinators or experienced school teachers. P4C school leads were appointed in all the schools in order to ensure the regular implementation of P4C sessions. The leads were also the point of contact for the evaluators to arrange for access to schools for the process evaluation.

The evaluation team at Durham University (and at Birmingham University when the trial started) was responsible for the research design, conducting the pupils' survey, data collection and data analyses, overseeing the completion of tests, and reporting of the independent evaluation.

Ethical review

The work was conducted in accordance with BERA's professional Code of Practice, and was approved by the University of Birmingham Ethical Review Board (the evaluators moved to Durham University after the trial had started).

The schools in this project were recruited by SAPERE. A Memorandum of Understanding was developed (see Appendix 1) by SAPERE in consultation with the evaluation team. All the schools willing to participate in the trial were obliged to sign the MOU which clearly stated the roles and responsibilities of the schools, SAPERE, and the evaluation team.

Opt-out consent forms were issued to parents of pupils in the participating schools before the schools were allocated in the treatment. The schools sent the opt-out consent letters to parents/carers informing them about the school's participation in the study. Parents were given the choice to opt out if they did not want their child's information to be shared for the research purpose or were not willing for their child to participate in the research. No withdrawal of consent from parents/carers was reported by schools.

The individuals' information was strictly anonymised in the final reporting and all data was coded and password protected. All participants in the interviews and observations were informed that participation was voluntary and that they could withdraw at any stage.

The school visits and observation of the P4C sessions were arranged with the school leads. The information achieved from these visits was anonymised, recorded in password-protected files, and was only used for the purpose of research.

Methodology

Trial design

The evaluation was a school-level randomised controlled trial with a waitlist control involving 48 schools across England. The number of schools was boosted from the original proposal in order to enhance the power of the trial. The revised plan was to recruit 50 schools, with more than half allocated to the control, but only 48 committed participants were signed up by the start of the trial. The randomisation process resulted in 22 schools allocated to receive P4C immediately, and another 26 to receive P4C after a complete calendar year (the process is described more fully below). This provided a cohort to act as a control and comparison group to assess the long-term effects of the intervention.

Pupils involved in the intervention were in Key Stage 2 (Years 3 to 6), and all formed part of the process evaluation. For the purpose of the impact evaluation, only pupils initially in Years 4 and 5 were involved in testing, with both groups contributing to the CAT results. In addition the initial Year 4 completed a survey of wider outcomes at the end of Year 5, and the initial Year 5 their KS2 results at the end of Year 6. The original Year 5 pupils in the control schools will provide a long-term counterfactual because they left primary school before the waitlist schools implemented P4C.

Schools

The 48 primary schools were recruited from five areas across England (Northeast, Northwest, Southeast, Southwest, and the Midlands) representing a range of geography, economy, local political control, population density, and levels of disadvantage. All schools have, or recently had, at least 25% of their pupils known to be eligible for free school meals (FSM). At least 10 of the schools had fewer than 60% of pupils achieving Level 4+ in English and maths, and with pupils making below-average progress in English and maths, in 2012 (or 2011). These were therefore disadvantaged schools.

Schools were approached both formally and informally, using existing contacts and local networks, meaning that they were focused in five areas rather than spread across England. A Memorandum of Understanding was sent to all schools prior to randomisation (Appendix 1).

Pupils

The evaluation followed Year 4 and Year 5 pupils. Opt-out consent forms were then sent by schools (not the evaluators) to parents to inform them of their child's involvement in the programme, outlining the purpose of the trial and the need to collect essential data while assuring them of confidentiality of potentially sensitive data.

Due to the difficulty of following the initial Year 6 pupils, the evaluation was not be able to assess their progress after at least one full year of progress or experience of the intervention. Schools were also loath to allow their preparation for KS2 to be disrupted. Year 6 pupils were therefore not assessed using CAT4s as part of the impact evaluation, but their involvement formed part of the process evaluation. All of this was as originally planned.

Intervention

Teacher training

Schools that undertook P4C received two types of training: introductory training for all staff in year 1 and additional days of in-school support as well as access to resources. All the training and support was provided by SAPERE-accredited trainers using standardised SAPERE methods.

All staff received two days of training (SAPERE Level 1). Some schools were able to devote two whole days to the training, while others decided to have one day plus an additional day of classroom demonstrations by the trainer together with two after-school sessions. The training included the following elements:

1. Explanations of the principles and methods of P4C
2. Demonstrations of P4C in action (with teachers or pupils)
3. An introduction to available resources
4. Advice on the evaluation of P4C
5. Advice on placing P4C in the school curriculum

In-school support

In addition to the introductory training, schools were offered additional support. The first step was for trainers to demonstrate P4C with children in each school. In subsequent sessions, the trainers supported teachers doing P4C by helping them plan lessons and giving them feedback. They also advised teachers who had taken on a leading role for P4C in the school. Nine support days were available to schools over the two years of the project. Most schools used all the support days but some did not, because of various in-school factors such as changes of priorities after inspections or changes of school management personnel.

Philosophy for Children in the classroom

This section describes the original idea, steps and procedures of the intervention. P4C is a whole-class intervention which aims to stimulate classroom dialogue in response to children's own questions about shared stories, films, and other stimuli. The classroom protocol must be adhered to, but it was up to schools how often they conducted P4C lessons.

The main emphasis of the intervention is to allow pupils to think and ask questions. With guidance from the teacher, the dialogue is focused not only on the chosen questions but also on the assumptions that lie behind the answers and the criteria used to make judgements. P4C aims to help pupils' to think logically, to voice their opinion, to use appropriate language in argumentation, and to listen to the views and opinions of others. The following are the main ten stages of a P4C session:

1. Getting set
2. Presentation of stimulus
3. Thinking time
4. Question making
5. Question airing
6. Question choosing
7. First thoughts
8. Building
9. Last thoughts
10. Review

1. Getting set

Pupils and teacher sit in a circle so everyone can see and hear one another. The teacher negotiates guidelines on conduct, and the aims of enquiry with children or reminds them of previously negotiated guidelines. Sometimes the guidelines are revised in the light of ongoing dialogue. Warm-up activities are sometimes used to aid co-operation, speaking, listening, and thinking.

2. Presentation of stimulus

The teacher introduces the planned material she has chosen in order to provoke pupils' interest, puzzle them or prompt their sense of what is important. Stimuli intended for P4C should reference the sort of 'big ideas' – such as truth, fairness, rights, knowledge, and friendship – that are likely to excite philosophical dialogue.

The range of stimuli used in P4C sessions includes short stories, poems, images, picture books, passages from novels, short video clips, newspaper articles, and material taken from other curriculum areas.

3. Thinking time

A minute of silent, individual thinking followed by pupils in pairs sharing interesting issues and themes, or jotting down key words. The teacher often records some of the key words and ideas that emerge.

4. Question making

The teacher may suggest a question based on the outcome of thinking time. This is appropriate if the children are very young or new to P4C. More commonly, the class is split into small groups and asked to decide on a question they think is interesting, worth discussing, and that requires an answer based on reasoned judgement.

5. Question airing

Children present their group's question so all can see and hear it. When all the questions are collected and recorded, children are invited to clarify, link, appreciate or evaluate the questions prior to choosing one for discussion.

6. Question choosing

When the listing of questions is complete, the next phase is to select a one as a dialogue starter. The selection is made by pupils using one of a range of voting methods. Depending on the age of the children and the level of their experience of doing P4C, the teacher may help them with their selection.

7. First thoughts

The teacher asks pupils to share their thoughts on the question with their thinking partner(s). The discussion floor is then open for all to share their views.

The teacher is moderator of the group so their role is to ensure that all pupils have a chance to speak. Confident and talkative pupils are likely to dominate the discussion but the teacher must encourage all pupils both to speak and to listen so that all may contribute to the 'community of enquiry'.

8. Building

Pupils participate in the discussion, building on other pupils' contributions, clarifying them, questioning them, and stating their own opinions. Whether agreeing or disagreeing the rule is to justify opinions with reasons. The teacher's role at this stage is to support pupils' reasoning, motivate them to question, and encourage them to take part in a dialogue with their peers.

Teachers will often prompt pupils to make moves such as imagining alternatives and consequences, seeking evidence, quantifying with expressions like 'all', 'some' or 'most', offering examples and counter examples, and questioning assumptions. Teachers promote and model a stance of 'fallibilism' – willingness to amend or abandon one's opinion in the face of a good argument to do so.

It is recommended in the P4C method to use some short gaps of silence or partner talk so that pupils can organise their thoughts and practise arguments with peers before sharing with the whole group. The teacher can also draw diagrams or make notes to keep track of significant arguments.

9. Last thoughts

The closing of the session involves last words from all pupils. Pupils might have the same opinion as they had at the beginning, or their view could have changed as a result of the dialogue. Pupils are invited to sum up their views concisely and without contradiction from others. They can sum up their views in a few words. This activity could either be a verbal statement or for a detailed reflection whereby a teacher could ask pupils to write a summary of their views.

10. Review of the session

The teacher invites reflective and evaluative comments about the enquiry with reference to broad criteria such as the guidelines the group has adopted (see stage 1). The teacher asks: 'What went well?' 'What could we improve on?' 'What do we need to do next?' The teacher could point to issues of pupils' behaviour and turn-taking in the session and ask them to reflect on their progress. The review could include suggestions on what else needs to be focused in the next P4C sessions.

Summary of Philosophy for Children intervention

P4C, as promoted by SAPERE, is a template to practise and organise a classroom session for philosophical enquiry. The intervention suggests useful stages of initiating, practising, and concluding the dialogue with pupils. However, it does not have any specified materials or stimuli that must be used; there are only examples and suggestions. The tens step outlined above are a guide to organising the classroom dialogue and can be used flexibly as the teacher's expertise grows. For example, the stages do not need to be completed all in one session. Choosing a question in one session and discussing it in another is a popular option.

Although there is no a set material or syllabus as such, the P4C website (www.P4C.com) provides useful teaching resources for teachers to refer to for topics, age-appropriate themes, sample resources, charts and tables to help categorise thinking points, and stimulus materials in different forms.

There is no special equipment required for this method except for usual material for teaching such as projector, board, pens, and sheets of papers. All pupils and teachers are required to sit comfortably in a circle facing each other for discussion. There is also the expectation for teachers to use existing curriculum material in their lessons when they judge it to have potential to stimulate philosophical discussion and to clarify key concepts in subject areas such as democracy, justice, nation, history, truth, evidence, beauty, art, belief, knowledge, tolerance, and theory.

Sample size

The sample consisted of 48 volunteer primary schools from Birmingham, London, Hull, Sheffield, Manchester, Hertfordshire, Staffordshire, and Stoke-on-Trent, thus including a wide variety of regions. SAPERE had planned to recruit 50 schools, but only 48 committed schools agreed within the time. It was felt that it was more important to have schools that were committed to the trial than to have a few more but run a real risk of school dropout. This sample was recruited in one phase by SAPERE who sent out an open invitation to the schools in the regions where SAPERE staff members were already working. This regional pattern may limit any attempted generalisation to all schools in England. All the schools in the study volunteered to participate. During the recruitment phase the schools were informed that this was a funded research trial and once recruited they would have to conform to the results of randomisation. According to the results 22 schools were to receive P4C treatment

immediately and the other 26 were to receive P4C after nearly two years (from September 2014). The imbalance in numbers was deliberate and is linked to the number of schools that SAPERE felt able to train in the first year. The training costs and costs for the support staff visits were covered for all 48 schools. SAPERE and the evaluators maintained a very good working relationship with all schools and completed the trial in terms of all schools providing results (there was no school dropout).

A total of 3159 pupils in Years 4 and 5 (entire year groups) took part in the trial, of which 1,550 were in the treatment group and 1,609 in the control group (Table 1).

Table 1: Number of schools and pupils in each arm of the trial

Groups	Schools	Pupils pre-tested (December 2012)	Pupils post-tested (January 2014)
Treatment	22	1550	1366
Control	26	1609	1455
Total	48	3159	2821

Power calculations make a number of assumptions that are not relevant here, but for illustration the estimate of sample size in the protocol was based on prior research evidence suggesting an effect size of 0.4. Assuming an intra-cluster correlation of 0.2 for the outcome scores, a minimum sample size of 480 pupils per arm would be needed (for 80% power to detect a difference of 0.4 with alpha of 5%) according to Lehr's formula (Gorard 2013). In fact, the situation is better than this, because of the correlation between pre- and post-tests scores for both Key Stage and CAT data. Thus a sample of 48 schools with over 3000 pupils should easily provide sufficient traditional 'power' to detect an effect in terms of CAT4 and KS2 progress outcomes.

Randomisation

Schools were randomised to two groups. A set of 48 random numbers was created, 22 representing intervention and 26 representing control schools, and then allocated to the list of participating schools in alphabetical sequence. Randomisation was conducted openly by the lead evaluator based on a list of schools and witnessed by colleagues.

Outcomes

The impact evaluation involved two main outcomes.

The first outcome was the KS2 results for all of the original Year 5 pupils, adjusted for their prior KS1 results in reading, writing, and maths. These were provided by the National Pupil Database (NPD) linked to unique pupil numbers (UPNs) supplied by all participating schools. The Department for Education matched the scores to the pupils for the evaluators, but would not permit these scores to be linked to some elements of pupil background data, or the CAT scores.

The second outcome was the gain score in the Cognitive Abilities Test (CAT4). This is a standardised assessment measure which is widely used to assess pupils' faculty of reasoning. GL Assessment has published the fourth edition of CAT for which age-related (age 7 to 17+ years) and online versions are available (http://www.gl-assessment.co.uk/products/CAT4-cognitive-abilities-test-fourth-edition). CAT4 as a standardised assessment tool was proposed by SAPERE as being appropriate for the kinds of measures that P4C might influence, and it had previously been used in the Topping and Trickey (2007) study. The online version was mutually agreed with the evaluators and EEF team members.

The trial included one cohort each of Year 4 and Year 5; therefore their age-related levels were used at the pre-test and post-test stages. The following sets were used for the relevant age and year groups:

- CAT4 A (Year 4 cohort 2012/13 pre-test) age 8 to 10
- CAT4 B (Year 5 cohort 2012/13 pre-test) age 9 to 11
- CAT4 B (Year 5 post-test 2013/14) age 9 to 11
- CAT4 C (Year 6 post-test 2013/14) age 10 to 12

CAT4 assesses four categories of reasoning skills:

- Verbal
- Non-Verbal
- Quantitative
- Spatial Ability

Each of the above categories includes items that assess mental processing such as identification, matching, determining links and series, classification, recognition, image retention, and analysis. These are deemed to be the core skills needed for critical thinking and other cognitive learning processes (Stein et al. 2013).

The four test sections are timed and adaptable to difficulty levels according to a pupil's early responses in the online test. The test provides detailed instructions and example and practice items. The online test is estimated to take around 50 minutes to complete.

Analysis

The intervention was offered to whole primary schools but the outcome measures were considered only for pupils initially in Years 4 and Year 5. For the initial Year 5 pupils, the primary outcome measure was their Key Stage 2 scores for reading, writing, and maths. For both year groups (initial Year 4 and initial Year 5) a further outcome measure was their post-test performance in the Cognitive Ability Test (CAT4) supplied by GL Assessment.

Further considerations were any changes in pupils' classroom learning and behaviour (e.g. pupils' relationship with school, teachers, and peers) and changes in teachers' behaviour (e.g. encouraging and demonstrating questioning and reasoning, less dominance in discussions). These were obtained from surveying pupils and teachers, interviewing them during the process evaluation, and observing the delivery of P4C sessions.

The schools administered CAT4 at two instances. An initial pre-test was carried out in December 2012 at the beginning of the intervention when there was equipoise. The post-test was taken in January 2014 by the initial Year 4 and initial Year 5 pupils. Since schools were no longer blind to allocation for the post-test, evaluators made sample visits to schools to observe the conduct of the test.

Every attempt was made to obtain complete test scores for all pupils even where they were initially absent or where they had left the schools during the trial. The post-test was arranged and set up in the destination schools of pupils who left during the period of intervention (23 pupils completed the test in the new destination schools). Every effort was made to ensure that every pupil was accounted for, even where this went beyond the two-week testing window. The results were analysed in terms of the original phase for each pupil (intention-to-treat).

The effectiveness or otherwise of P4C is represented by:

- the effect size (Hedges' g) for the standardised gain score from KS1 to KS2 in reading, writing, and maths.
- the effect size (Hedges' g) for the standardised gain score from CAT4 pre-test (CATA for Year 4 and CATB for Year 5) to post-test (CATB for Year 5 and CATC for Year 6).
- the effect size (Hedges' g) for the standardised gain score from CAT4 pre-test (CATA for Year 4 and CATB for Year 5) to post-test (CATB for Year 5 and CATC for Year 6) for each of the test subscales – verbal, quantitative, non-verbal, and spatial.

Additional analyses were performed for each year group, repeating the overall analysis but using scores for only those pupils eligible for FSM (as pre-specified). Other analysis, which were not pre-specified, compared boys and girls separately, and those above or below the median (middle) score in the CAT pre-test.

Process evaluation methods

The process evaluation provided formative evidence on all phases and aspects of the intervention from the selection and retention of schools, through the training of teachers, to testing the eventual outcomes. This was used to assess fidelity to treatment, implementation issues, and the perceptions of participants, including any resentment or resistance.

The evaluation team made 30 visits to treatment schools, usually one at the beginning of the intervention and one towards the end to observe changes in teacher and pupil behaviour. Schools were visited repeatedly to assess progress. The trips included observations of the initial training of teachers as well as the delivery of the programme in the classroom. Evaluators attended three training sessions as participant observers, noting the process of implementing P4C, the methods of delivery, and also teachers' responses to the training. The observations of P4C in action were non-intrusive, with the evaluator sitting either inconspicuously at the back of the classroom or more usually as part of a circle but not taking part in the dialogue unless directly addressed. Interviews with teachers and pupils were also conducted during these visits. These interviews were very informal conversations with teachers and pupils who were involved in doing P4C intervention. In each visit a prior meeting was set up between the P4C lead and the teaching staff to discuss the lesson to be taught that day. The evaluation team members also observed the debriefing sessions after lessons in order obtain teachers' feedback on P4C sessions.

The aims of the observations and interviews were to help answer the following questions:

1. Is the suggested number of sessions adhered to?
2. Are children doing P4C sharing their ideas more with each other in a critical but friendly way?
3. Are questioning and reasoning being prompted and demonstrated in lessons?
4. Are instances of questioning and reasoning increasing?
5. Is there less dominance by the teacher in discussions?
6. Are children taking more responsibility for the questioning and reasoning?
7. Are teachers and children talking about significant concepts?
8. Are teachers' perceptions of children changing?
9. Are teachers' perceptions of their own work changing?
10. Are children's perceptions of themselves and school changing?

The above set of questions was used as a guideline during school visits and P4C observation sessions. The details in response to these questions have been described in the results sections of the process evaluation.

Eight schools were visited for in-depth observations of the P4C sessions. The schools selected were the first that responded to the call of independent evaluators' observations. The evaluation team members observed eight session of P4C being conducted in different schools. The process of

delivery was observed by the evaluation team members, keeping the above-mentioned guidelines in mind. After the observations pupils involved in the sessions were generally asked their opinions of P4C. This generally included questions about their level of enjoyment during the sessions, aspects of P4C they liked or disliked, changes they perceived in their level of confidence, friendship with peers, and the quality of communication with teachers. That said, there was no pre-planned structure of these conversations and the purpose was to gather a general overview of pupils' attitudes to P4C.

Observations were first recorded as handwritten field notes by the evaluation team member who conducted the school visit. After each visit the evaluator team member developed a report of the visit which included the detailed descriptions of the field notes, teachers' feedback and comments and details about conversations with the pupils. All three evaluators conducted visits and enough observations were accumulated to develop an overall understanding of the process as practised, teachers' and pupils' views and feedback on P4C, and challenges in the implementation process.

The implementation of P4C in the schools was closely monitored by SAPERE to ensure that the delivery adhered to the protocol. A P4C-accredited trainer provided regular feedback reports to SAPERE about the quality and the level of implementation in the schools ('accredited' is the term used by SAPERE to refer to its trainers for each school and indicates a high level of expertise and experience). These reports provided insights on the barriers and challenges in implementation. Each school was given a score based on frequency of lessons, and observed adherence to the protocols.

The evaluators observed the conduct of the CAT4 tests and the survey in a sample of six schools each. These observations were carried out to understand the process of testing, and to flag any issues that might arise with online testing. The school staff members responsible for conducting the CAT4 pre- and post-tests were not the P4C teachers. This was a measure taken to ensure that P4C pupils or the control pupils did not receive preferential treatment during the test since the teachers were no longer blind to allocation.

Impact evaluation

Table 2: Timeline

Dates	Activities
September–November 2012	School recruitment by SAPERE, preparation meetings between EEF, SAPERE and evaluators
December 2012	CAT4 pre-test conducted with Year 4 and 5 cohorts in all schools, randomisation carried out by evaluator, group allocation provided to SAPERE and school leaders
January–March 2013	P4C teacher training workshops conducted for all teachers in 22 schools in treatment group
February–December 2013	P4C implementation in treatment schools, SAPERE support staff visits to treatment schools, evaluation team visits to (all?) schools
January–February 2014	CAT4 post-test conducted in all schools for all initial Year 4 and 5 pupils. Pupil survey conducted in all schools for the initial Year 4 pupils
March–April 2014	Information completed on pupils missing in the sample. The missing pupils in the sample were followed to their new schools
June–December 2014	Procured data from National Pupil Database on KS2 results for the original Year 5 pupils, data analysis, and report writing. Control schools permitted to start the intervention

Participants

Schools

Schools were recruited by SAPERE. A total of 48 schools were recruited (Table 3) and 22 of these were randomised to the first phase treatment group to start P4C in January 2013, and the remaining 26 schools to the second phase forming the control group. None of these 48 schools had prior experience of using P4C.

Table 3: Number of sample schools by type, denomination, and most recent OFSTED result

School category		Denomination		OFSTED effectiveness	
Academy	35	**Church of England**	8	**Outstanding**	5
Community	1	**Roman Catholic**	7	**Good**	22
Voluntary aided	9	**Does not apply**	33	**Requires improvement**	18
Foundation	3	–		**Inadequate**	3

All schools were primary schools, several with relatively high overall FSM eligibility (over 50%), mixed gender, and all had a diverse pupil ethnic group (see pupil characteristics, below). No school dropped out of the study so there is no attrition at school level. All schools were volunteers from areas in which SAPERE had existing trainers, and this might affect any future generalisation of results.

Pupils

The intervention was implemented using whole year groups from Year 3 to Year 6. The evaluation included pupils in Year 4 and Year 5. The headline outcome measure for KS1 to KS2 progress includes all of the original Year 5 pupils in all 48 schools (1529) with figures available from the National Pupil Database (NPD). There is no evidence of attrition. The evaluators were not permitted by DfE to link these attainment records to the UPNs of pupils who completed the CAT. This has no implications for the long-term comparison, since all future assessments of the impact of this intervention can compare the subsequent results of all pupils initially in Year 5 in all treatment and control schools. For the CAT scores, Years 4 and 5 initially had 3159 pupils at pre-test. In the final

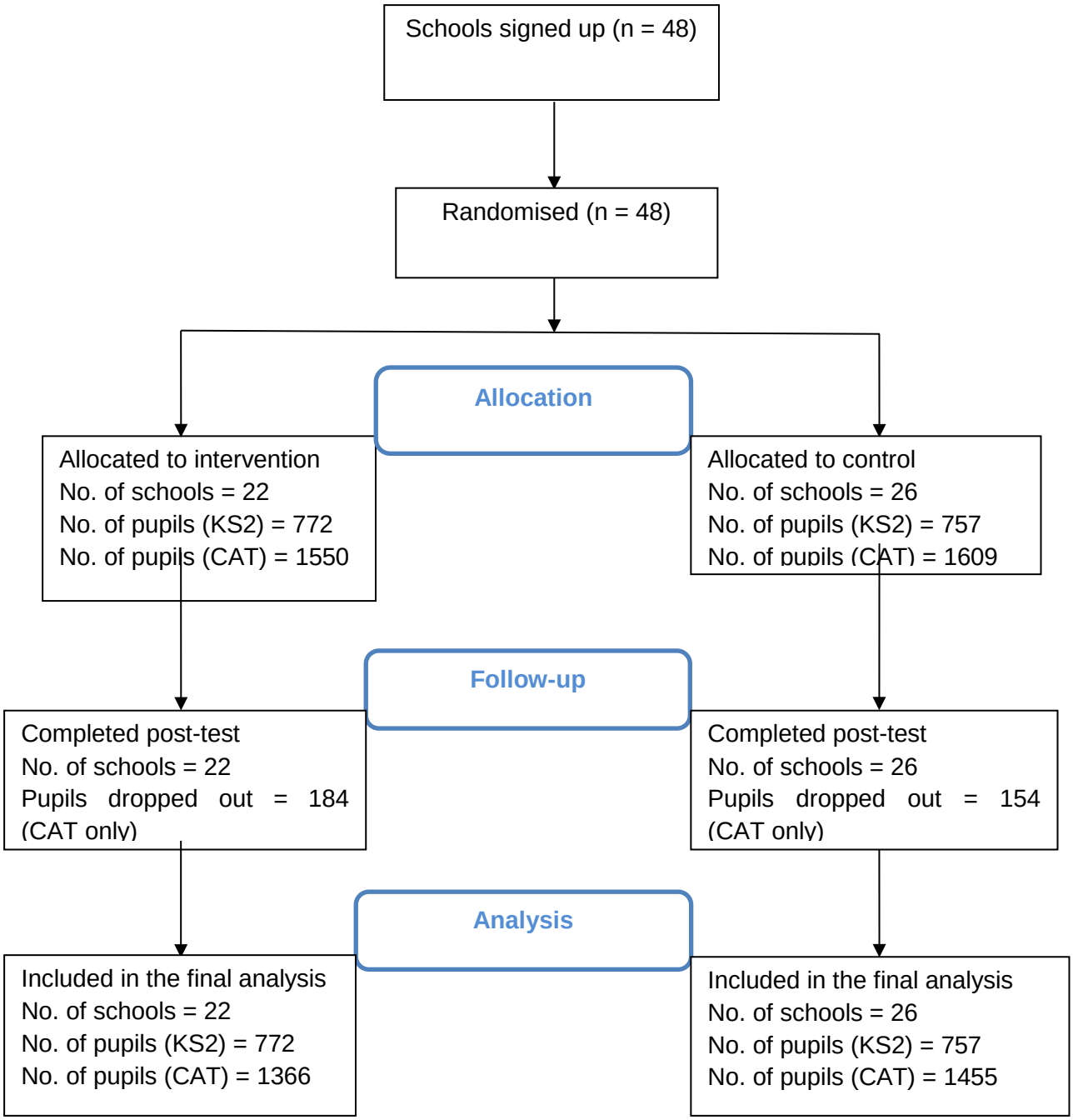
analysis 338 pupils did not provide post-test data and could not be followed for the post-test due to reasons such as: left country, started home schooling, from traveller group, did not attend the school after the first few days. Another difficulty was that several of the schools would not provide details of the destination school. The evaluation team along with the school leads tried to approach all the missing pupils to complete the post-test. There were 23 pupils who completed the test as a result of this follow-up. The attrition for CAT scores is 10% of the total pupil numbers in the year groups and schools selected for the study.

The overall sample was relatively disadvantaged (Table 4). The two groups were well-balanced in terms of sex, FSM eligibility, and SEN status. The relatively small number of children speaking English as an additional language was disproportionately represented in the control (65%). The prior attainment and initial CAT scores are described later in the report.

Table 4: Balance between intervention and control group: percentage of pupils with background characteristics in each group (where these are available)

	Intervention %	Control %	N
Male	51	52	1452
Female	49	48	1369
FSM	48	46	1478
SEN	18	19	515
EAL	9	15	378
Non-White UK	31	23	759

Figure 1: Participant flow chart



Outcomes and analysis

Attainment

The Key Stage 2 results were obtained for Year 6 (2013/2014). The total of 1529 pupils was included in this analysis, representing the available scores for all original Year 5 pupils in all schools. The effect sizes were determined for reading, writing, and maths.

At the outset the treatment and control groups were reasonably balanced, with the control group having slightly better KS1 scores in reading, writing, and maths (Tables 5 to 7). By the end the treatment group had narrowed this gap in all three subjects, especially for KS2 scores in reading and maths. For this reason, the key stage results are all presented as gain scores representing progress from KS1 to KS2. Because the fine point scores on the NPD use a different metric for KS1 and KS2, each set of scores is converted into a standardised format (z-scores) before the gain score is computed.

Table 5: KS1 to KS2 Reading progress

	N	Mean KS1 points z-score	SD	Mean KS2 fine points z-score	SD	Gain z-score	SD	'Effect' size
Treatment	772	-0.08	1.01	-0.02	1.01	+0.06	0.88	+0.12
Control	757	+0.08	0.98	+0.02	0.99	-0.05	0.91	-
Total	1529	0	1	0	1	0	0.90	-

Table 6: KS1 to KS2 Writing/GPS progress

	N	Mean KS1 points z-score	SD	Mean KS2 fine points z-score	SD	Gain z-score	SD	'Effect' size
Treatment	772	-0.07	1.03	-0.05	1.00	+0.01	0.77	+0.03
Control	757	+0.07	0.96	+0.06	1.00	-0.02	0.90	-
Total	1529	0	1	0	1	0	0.84	-

Table 7: KS1 to KS2 Maths progress

	N	Mean KS1 points z-score	SD	Mean KS2 fine points z-score	SD	Gain z-score	SD	'Effect' size
Treatment	772	-0.09	1.04	-0.04	1.01	+0.04	0.74	+0.10
Control	757	+0.08	0.95	+0.04	0.99	-0.04	0.82	-
Total	1529	0	1	0	1	0	0.78	-

On viewing the gain scores there is evidence that P4C may have a positive impact on pupil attainment at KS2, equivalent to about two months' extra progress for reading and maths, after just over a year of implementation. There is no clear benefit for writing in the overall results, which is perhaps not surprising since there is no writing element in P4C; P4C involves reading and reasoning.

The results in Tables 5 and 7 are unlikely to be due to chance, and no dropout was recorded from NPD or reported by DfE when they matched the data. Tables 8 to 10 show the results (pre-test, post-test, and gain scores) for only those pupils known to be eligible for free school meals (FSM). The 'effect' sizes are even more positive than for Tables 5 to 7. These results do not have the power of a trial, since the FSM cases were not randomised as such. But they do suggest that P4C is at least as

effective for FSM pupils, and that P4C could be one way of reducing the poverty gradient in Key Stage 2 results.

Table 8: KS1 to KS2 Reading progress: FSM-eligible pupils only

	N	Mean KS1 points z-score	SD	Mean KS2 fine points z-score	SD	Gain	SD	'Effect' size
Treatment	265	-0.40	1.02	-0.16	1.00	+0.24	0.92	+0.29
Control	233	-0.10	1.01	-0.12	1.06	-0.02	0.87	-
Total	498	-0.26	1.02	-0.14	1.03	+0.12	0.91	-

Table 9: KS1 to KS2 Writing/GPS progress: FSM-eligible pupils only

	N	Mean KS1 points z-score	SD	Mean KS2 fine points z-score	SD	Gain	SD	'Effect' size
Treatment	265	-0.36	1.05	-0.25	1.00	+0.12	0.80	+0.17
Control	233	-0.10	0.98	-0.12	1.03	-0.02	0.85	-
Total	498	-0.24	1.02	-0.19	1.01	+0.05	0.82	-

Table 10: KS1 to KS2 maths progress: FSM-eligible pupils only

	N	Mean KS1 points z-score	SD	Mean KS2 fine points z-score	SD	Gain	SD	'Effect' size
Treatment	265	-0.36	1.10	-0.28	0.93	+0.09	0.80	+0.20
Control	233	-0.03	0.95	-0.11	1.05	-0.08	0.91	-
Total	498	-0.21	1.04	-0.20	0.99	+0.01	0.86	-

CAT4 Results

The two groups were reasonably well-balanced in terms of CAT scores at the outset, but again with the control slightly ahead (Table 11). For this reason the results are generally presented as simple gain scores from pre-test CAT to post-test CAT. Overall, the treatment group made a slightly larger gain in CAT scores than the control (around one month's extra progress in just over a year). Clearly, P4C is doing no harm to pupils' attainment or cognitive abilities. The results are too small, given the inevitable vagaries of such a study including some attrition for the pupils providing CAT scores, to state that these gains are definitely the result of P4C, but overall there is some promise here.

Table 11: Overall CAT4 gain score

	N	Pre-CAT4	Standard deviation	Post-CAT4	Standard deviation	Gain score	Standard deviation	'Effect' size
P4C	1366	94.37	11.24	96.59	12.26	2.22	7.59	+0.07
Control	1455	95.20	11.19	96.90	11.90	1.70	7.32	-
Total	2821	94.80	11.22	96.75	12.07	1.95	7.46	-

Tables 12 to 15 show the same kind of gain scores for each of the four subscales of the CAT test – verbal, quantitative, nonverbal, and spatial. P4C had the biggest impact, on average, in terms of the verbal subscale. This is both to be expected and ties in with the greater gain for the treatment group in KS2 reading.

Table 12: Gain scores in verbal subscale of CAT4 (B)

	N	Gain score	Standard deviation	'Effect' size
P4C	1301	+2.70	15.17	+0.08
Control	1417	+1.60	10.79	-
Total	2718	+2.13	13.08	-

Note: the number of pupils differs in Tables 13 to 16 because some pupils did not complete all sections of the test.

Table 13: Gain scores in quantitative subscale of CAT4 (B)

	N	Gain score	Standard deviation	'Effect' size
P4C	1128	+0.14	24.22	-0.01
Control	1355	+0.34	12.32	-
Total	2483	+0.25	18.69	-

Table 14: Gain scores in non-verbal subscale of CAT4 (B)

	N	Gain score	Standard deviation	'Effect' size
P4C	1367	+3.96	23.72	+0.04
Control	1451	+3.29	12.32	-
Total	2818	+3.62	18.74	-

Table 15: Gain scores in spatial subscale of CAT4 (B)

	N	Gain score	Standard deviation	'Effect' size
P4C	1132	+2.83	27.31	+0.07
Control	1351	+1.35	12.94	-
Total	2483	+2.03	20.77	-

All of the overall gains in CAT scores come from the pupils who were in Year 5 initially (Tables 16 and 17). There is no difference between the groups for the younger cohort and a small but noticeable 'effect' size for the older cohort.

Table 16: Overall CAT4 (B) gain score, Year 4 initially

	N	Gain score	Standard deviation	'Effect' size
P4C	691	1.46	7.30	-0.01
Control	758	1.55	7.41	-
Total	1449	1.51	7.35	-

Table 17: Overall CAT4 (C) gain score, Year 5 initially

	N	Gain score	Standard deviation	'Effect' size
P4C	675	2.89	7.82	+0.14
Control	697	1.85	7.22	-
Total	1372	2.36	7.54	-

A similar bifurcation appears in the scores for FSM-eligible and non-FSM pupils (Tables 18 and 19, which was pre-specified analysis). Pupils who are eligible for FSM have shown no gain from using

P4C and the overall result is explained solely by the small but noticeable 'effect' size for pupils not eligible for FSM. This is in sharp distinction to the results for KS2 attainment (see above).

Table 18: FSM pupils CAT4 gain score

	N	Gain score	Standard deviation	'Effect' size
P4C	697	1.45	7.17	-0.02
Control	781	1.66	7.36	-
Total	1478	1.56	7.27	-

Table 19: Non-FSM pupils CAT4 gain score

	N	Gain score	Standard deviation	'Effect' size
P4C	669	2.92	7.94	+0.15
Control	727	1.77	7.24	-
Total	1360	2.34	7.62	-

Other subgroup analyses, which were not pre-specified and do not have the rigour of the overall trial, are also of interest in terms of the ongoing process of understanding the results. These results are presented in Appendix 2. They show that it was pupils in the higher-scoring half of CAT scores at the outset who are creating the overall positive result (Tables 18 and 19). The lower-scoring half of the pupils showed no gain in CAT scores. These results are suggestive that benefits from P4C may be restricted to the oldest cohort (also the pupils with the reported gain in KS attainment), who are less disadvantaged and higher-scoring at the outset.

For completeness it is worth also reporting that the CAT gain scores were similar for boys and girls, lower for SEN (special educational needs) pupils and much higher for pupils from an ethnic minority or for whom English is an additional language (see Tables 22 to 27 in Appendix 2). As above, these analyses were not pre-specified, and are presented here for interest and completeness. None has the strength of a trial. The equivalent results for KS2 are not available.

Cost

SAPERE provides professional teacher training workshops and courses in adopting P4C. There are three levels for a teacher to become an expert P4C practitioner. Level 1 foundation training is the basic teacher training level, after which teachers can start implementing P4C with their groups. Subsequent training levels build advanced P4C skills. For the current research project Level 1 training was provided to all the teachers taking part.

After a P4C foundation training a primary school can implement the intervention at a whole-school level. Table 20 shows that the per-pupil cost of this intervention was roughly £16 for a primary school where 240 pupils are enrolled, not including any staff cover. For the evaluation, these costs were met by EEF both for the treatment schools and the waitlist control schools. SAPERE states that its current programmes typically cost £25–30 per pupil as they do not benefit from some of the economies of scale in the EEF project.

Table 20: Cost components of the P4C intervention

	Cost	Time/days
Level 1 training (for a whole group of teachers which include 6 to 8 staff members per school) and expert staff support costs	£1,300	2 days
Support days	£300×8 = £2,400	8 days
SAPERE annual membership	£30	12 months
P4C website subscription	£10	12 months
P4C stimulus books	£200	One-off cost
Total cost for one primary school	£3,940	12 months

Process evaluation

Implementation

It was up to schools how often they conducted SAPERE's P4C lessons. Usually schools implemented one P4C lesson per week in place of the usual literacy session. A few faith-based schools used religious studies sessions instead, and some schools had more than one session per week. As there is no prescribed syllabus in P4C this approach could have been adopted in other lessons such as English, maths, PSHE, history, or geography. However, the teachers reported that their regular lessons have fixed syllabi and set targets to achieve and it was difficult to follow the P4C format in the regular lessons. P4C does not directly teach elements of the National Curriculum measured through SATs and it was reported as a challenge to make space for P4C in the regular teaching schedules. Teachers used classrooms, assembly halls, and libraries as the venue for conducting P4C sessions.

According to feedback from teachers who were asked about the challenges of implementing P4C, it was understood that the success of the intervention depended on incorporating P4C into the timetable on a regular basis, and of making it part of normal school interaction. As this intervention does not target a specific subject, the commitment of staff and school management is required in order to embed the practice of P4C in the school culture. In addition, the implementation needs to be monitored. Otherwise, when teachers face consequences for not succeeding with initiatives that are closely monitored, P4C will tend to be 'crowded out'. According to interviews with the teachers and school leads, successful P4C requires good preparation of ideas and resources before they are presented to any pupil enquiry group.

Teachers' reflections on the training suggested that training was essential because the intervention is based on exploring concepts rather than hands-on activities or delivering information or skills. They reported that without attending P4C training the process of intervention could not have been implemented as per the protocol of P4C. Teachers could have various styles and interpretation of conducting P4C without the training. The training covered a broad range of concepts that could be used in the sessions in different ways. Conceptual exploration through P4C was supported through the website where a wide range of resources and ideas are available.

The teachers who conduct P4C need to be aware of their own biases and beliefs that could influence pupils' involvement and learning process. As observed by the evaluators during the sessions, the pupils often only shared their views once they had developed trust in the teacher and were confident that their views were equally important and respected and could be voiced without retaliation from teachers and peers. The evaluators were informed by pupils' feedback that during P4C they were allowed to share and question without being interrupted by the teacher. Some pupils also reported that they felt more relaxed talking during P4C as compared to normal lessons because they knew that the teacher would not discourage them from talking and discussing.

The evaluators observed that P4C sessions tended to improve pupils' confidence and engagement when teachers were equal participants in the enquiry circle with the pupils, rather than taking their normal position of authority in the classroom. In some sessions observed at the beginning of the project, teacher talk time was more dominant than pupils' participation. It was observed that the teachers needed feedback or practice and time to negotiate their participation level in the sessions and to let pupils talk and discuss more. However, in observations made in the later period it was noticed that the same teachers moderated sessions which were well balanced in terms of their own input and pupils' participation.

It was observed by the evaluators that a complete P4C session should usually cover the ten steps of enquiry, otherwise pupils would not gain the sense and purpose of the whole activity. For example, in one of the sessions the discussions initiated were not summed up nor sufficiently reviewed. The

session was rushed to the end as the time for the session was passing quickly. As a result, one observer noted that some pupils had not really understood the sense and purpose of the discussion because there was no proper conclusion. Sometimes pupils said that the questions were not fairly selected and pupils cheated and voted for their friends' questions. It was observed that if questions were not fairly selected through voting then possibly pupils would miss the chance of learning the process of fairness. In one of the sessions the pupils were not given enough thinking time and this was possibly the reason that they could not reflect on the issues for developing interesting questions.

The control schools were on the waitlist to receive P4C teachers' training and implement the intervention in the schools. The developers and evaluators ensured that none of these schools used P4C during the period of the trial. The control condition was therefore 'business as usual'.

Barriers to delivery

P4C is a popular intervention worldwide though materials and methods vary. The SAPERE model of P4C is not expensive to set up and, once fully embedded in a school, can be sustained without developer involvement. However, there are some clear challenges to the delivery and implementation of P4C.

The main challenge reported by teachers and school leaders was the difficulty of embedding P4C in the fully-packed timetable and with targets for literacy and numeracy from the National Curriculum. Teachers reported that there is often not enough time to be regularly devoted to P4C when there are so many other activities going on. P4C school leaders reported that the teachers do not see this intervention as easily fitting with the goals of subject-based teaching. P4C is particularly focused on underlying key concepts such as 'knowledge' and 'belief'. Deep discussion of these kinds of foundational concepts is often seen as not being as important to subject teaching as the learning of subject content.

P4C is a practice of dialogic teaching. There is no complete syllabus or unyielding methodology for the SAPERE approach to P4C. Without clear guidance or set discussion topics, there is a danger that this approach may be open to the influence of teachers' biases, beliefs, and ideologies, and examples of this were noted in our fieldwork.

A few pupils in some of the necessarily large enquiry groups were sometimes neglected by the teachers and their peers. It was observed in the sessions and was also reported by the pupils that they wanted to contribute at certain points and put their hand forward but teachers just moved on or gave the opportunity to another pupil. Where, as is desirable, the speaker decides who speaks next there is a fine line between a genuine back and forth between two pupils necessary for sustained argument, and abuse of the system by groups of friends.

Is the intervention attractive to stakeholders?

Schools/Heads

The intervention is appealing to many schools as a way of raising and debating pupil/school discipline problems in an enquiry group. The school leads reported that they discussed the concepts of bullying, racism, lying and cheating, equality and fairness which are core issues of school discipline and ethos. P4C was reported by the teachers to be very helpful in pupils thinking critically about these issues, raising questions, reflecting on their experiences, and coming to fair conclusions. P4C creates an opportunity for school leads to engage with pupils and develop a whole-school culture of thinking, listening, speaking, and arguing. Some of the examples of questions discussed in P4C observed sessions were as follows:

- Should a healthy heart be donated to a person who has not looked after themselves?
- Is it acceptable for people to wear their religious symbols at work places?
- What is honest feedback?
- Why do men receive more sponsorship than women in tennis?
- Can you and should you stop free thought?
- Is it OK to deprive someone of their freedom?
- What is bravery?

The above list of questions was created by pupils themselves from the given stimuli such as a story or short video, using a blind voting system. The substance of these questions is clearly relevant to the broader purpose of schools.

Pupils

The pupils who were interviewed generally showed their appreciation of the P4C sessions. The activity gives control to pupils in developing questions and voting for the questions. Pupils enjoyed the feeling of being in charge of the process. Several pupils in different schools reported that they got to know what their peers think during P4C, which may be difficult in other classroom situations. Older pupils reported solving their grievances with their peers during P4C sessions. A pupil commented that the children fight less in the playground because they had improved the way they talk. All these details on pupils' experiences were based on informal conversations with pupils.

The most common reaction reported by pupils was that they liked and enjoyed the idea of generating questions and the openness of asking a wide variety of questions. Some of the older pupils said that it was hard for them to develop questions in the beginning because they had never done anything before where they were asked to create questions. The pupils felt P4C was a liberating experience in terms of asking, sharing, and arguing. One of the pupils said:

> *I found creating questions difficult. It was hard. I didn't like it in the beginning. I have become better now. I have learned it quickly.*

Another pupil said:

> *I like one thing about P4C that there is no question right or wrong. All we think can be said and we listen also everything.*

Some pupils said that sometimes the topics were boring, especially if they are commonly discussed in lessons or elsewhere. Pupils wanted exciting stimuli and new concepts to be explored in every P4C session. Some pupils also would have preferred spending P4C time doing activities like sport, while a small number would have preferred a 'normal' lesson.

Teachers

Most of the teachers reported that the time constraints and other priorities in the curriculum often made them neglect P4C. Many said that preparation for the sessions demanded a good deal of teachers' time, although it is not clear whether this is because it was new and therefore additional. In the interviews all teachers reported that they enjoyed doing P4C and that it improved relationships with their pupils. Some teachers also reported surprising changes in some pupils' behaviour. During P4C some of the low-achieving and quiet pupils started gaining confidence through participation. Teachers also reported some indirect and positive influence of P4C on pupils' performance in English. One teacher remarked:

> *I feel much more comfortable listening to the children and allowing them to share ideas and have a more open classroom environment. Children are much more willing to listen to each other and are able to articulate their ideas towards each other*

Another teacher commented:

> *P4C has a huge impact on speaking and listening and building on each other's ideas.*

The intervention may be attractive to teachers for several reasons. There is a lot of teaching material available on the P4C and SAPERE websites and the teacher training is followed up by P4C-trained staff visits to give feedback to teachers doing P4C in real classrooms. P4C does not prescribe a specific syllabus, therefore teachers have freedom to adapt this intervention. There is no specific pupil grouping or required group size for a P4C session. It can be a whole-class intervention but teachers are free to organise pupil subgroups as the need arises.

Perceived outcomes

Teachers

Teachers often reported that P4C had a positive influence on the wider outcomes such as pupils' confidence to speak, patience when listening to others, self-esteem, well-being, and happiness. Some also believed that it had a positive influence on pupils' sense of judgement. A few comments from teachers are:

> *It has been fascinating to see children who are usually quieter or more reticent developing their thinking and becoming more confident.*

> *It gives children the confidence to know that they can talk in class discussions and that their viewpoint is a valuable one.*

Other teachers said:

> *Children are better at taking turns and listening to each other particularly when working in groups*.

> *They aren't content to just know surface knowledge about subjects. They always have questions and want to look into topics more deeply.*

Another commonly perceived outcome by teachers was that the disadvantaged pupils gained more from P4C than the high-achievers because it does not directly target academic performance. Pupils with English as an additional language were perceived to have gained more vocabulary in P4C through listening to others and participating in the discussion. Pupils with behavioural problems were perceived to have learned ways to control their behaviour and reason their way through problems. These are teachers' own perceptions of P4C's impact and could differ from evidence obtained in the evaluation.

Teachers believed that the impact of P4C extended across the curriculum. They believed that pupils who were doing P4C were increasingly becoming engaged in the classroom and were asking more questions in all lessons.

The teachers also reported that doing P4C had improved their own teaching style. P4C teachers adopt the style of facilitator in the classroom and gradually give more freedom to pupils to create their own questions rather than just answering leading questions asked by teachers. A teacher said that by doing P4C she became more confident in giving some autonomy of learning to her pupils rather than being more authoritative and controlling what the pupils would do. Other teacher comments reflected similar views:

The children are more comfortable to question me so the classroom has become a more collaborative learning environment.

Since beginning P4C a noticeable change has been seen in their questioning and enquiry skills across all subjects and I have become more aware of my own questioning of them.

I am now seeing a change in the way I respond to pupil discussions and my teaching now involves much more speaking and listening and encouraging pupil opinions

Pupils

The pupils often felt that P4C helped them in generating new opinions and learning new words. Some of the pupils felt that P4C improved their ability to communicate during group tasks and in other classroom activities. Pupils reported a general decline in classroom noise, which they attributed to waiting skills developed during P4C sessions. Some pupils also reported developing better relationship with their peers because they felt that they knew more about each other than they knew before due to P4C sessions. One of the pupils said:

I never had talked much with Adam. We were kind of strangers from each other. I know him now he gives good points when we do P4C. I like to be his partner in making questions.

Fidelity

The fidelity to the protocol of the intervention in the treatment schools was judged in two ways: the regularity of implementation and the quality of the sessions. SAPERE-accredited trainers visited schools and provided feedback on the regularity and quality of P4C sessions (this process itself was not part of the work done by the evaluators, who were merely provided with the resulting judgements). SAPERE did this through informal reports and also by completing an online survey where they used their own judgement to score their schools on the quality of the P4C sessions. According to the trainers' assessment of the schools the following details were achieved (Table 21).

Table 21: Combined 'scores' for regularity and quality of intervention (out of 16)

Number of schools	Maximum score on the quality and regularity of P4C sessions	Remarks
2	15	Excellent implementation
5	14	Implementation nearly at the highest satisfaction level of the accredited trainers
7	13	Implementation at the satisfaction level of the accredited trainers
2	11	Implementation needs improvement
6	10 and below	Implementation clearly not at the satisfaction level of the accredited trainers

Six schools did not implement P4C to the satisfaction of the expert staff (SAPERE trainers who provided extra support visits and demonstration in the schools), and a further two schools had deficiencies either in substance or frequency. These schools faced various challenges during the year of the trial. These difficulties included: a new headteacher who had no awareness of school involvement in the project, P4C lead teachers who left their school without providing a proper briefing for the new staff member, P4C trained staff frequently changing jobs, and in some cases poor OFSTED results that changed the nature and priorities of the school practices. There were also managerial-level transitions within schools that took place during the period of the trial and that affected the implementation of the intervention. SAPERE trainers gave these schools follow-up support and modelling of P4C sessions. To some extent these schools continued conducting the intervention but could not achieve the performance levels expected by SAPERE-accredited trainers.

Some schools did not actually implement P4C from the outset, but used it only in certain lessons such as religious studies.

Using these school scores for fidelity and correlating them with the attainment gain scores, there is no discernible relationship (Pearson R correlation of +0.02 for CAT scores, -0.01 for reading, -0.10 for writing, and -0.07 for maths). Schools judged to be implementing P4C better did not show greater overall gains than those implementing it poorly. This suggests that the judgements were not accurate or that if there is any impact it comes from implementing P4C at all more than precisely how it is implemented.

Formative findings

The formative findings emerged from the process evaluation. This involved gathering teachers' views, pupils' feedback, observations of the sessions in the beginning and near the end of the trial period, and expert staff members' assessment of the schools.

Integrating P4C in the school curriculum

The main challenge of implementing P4C is that it does not directly address schools' targets in literacy and numeracy targets. Schools can integrate P4C into existing subjects such as literacy or create timetable space for P4C as an individual activity.

The problem of integrating P4C in the school curriculum can be addressed by setting some targets to achieve. This could be in various forms, such as completing simple tests on critical thinking, or developing a project. This could lead pupils to obtain some sense of accomplishment in doing P4C.

Progress measures in P4C

The intervention just describes the aims and process to be followed. At the time of the evaluation, there were no clear P4C progress indicators that could explain individual and class performance to the pupils, teachers, and parents (although SAPERE has since developed these).

Teachers' training and guidance

Teachers were free to select topics and materials for the P4C sessions. The P4C website gives details on appropriateness of the content and age relevance of the materials and topics. This means that teachers who do not subscribe to the website have to make their own selection of materials and use their judgement about the appropriateness of the materials.

As variations of P4C are used around the world, it can be hard to pin down the content. However, there are some specific guidelines for the teachers to follow with regard to the selection of material and the discussion in the enquiry sessions.

The suitability of CAT4 relevance for P4C

Although P4C aims to improve pupils' cognitive abilities to reason, there are few precise reasoning skills that are targeted in the method. The P4C approach is broad, flexible, and based on dialogic practice, with a focus on developing a style of reasoning and the vocabulary underpinning common arguments. However, CAT4 measures specific skills such as verbal and non-verbal reasoning, quantitative judgements, and spatial ability. P4C aims to develop reasoning and other cognitive abilities through conceptual enquiries and dialogue on the issues raised by the stimulus material, while CAT4 is about mental exercises through shapes, patterns, sizes, signs, numbers, and word groups. CAT4 was used, as suggested by SAPERE, because it had been used in previous smaller 'successful' trials. As here, it is important that the test is independent of the intervention, but SAPERE were happy that it was a fair test of what the intervention could influence.

Control group activity

The control schools (on the waitlist) were to receive P4C teacher training and implement the intervention after the trial is completed. The evaluators ensured that none of these schools used P4C during the period of the trial. However, there were several similar approaches such as Thinking Hats or Circle Time that target critical thinking skills and they could have been used in these schools. There is no ideal control situation but the evaluators have visited control schools since the end of the intervention and are not aware of any systematic approach to critical thinking adopted in the control schools.

Conclusion

Key conclusions
1. There is evidence that P4C had a positive impact on Key Stage 2 attainment. Overall, pupils using the approach made approximately two additional months' progress on reading and maths.
2. Results suggest that P4C had the biggest positive impact on Key Stage 2 results among disadvantaged pupils (those eligible for free school meals).
3. Analyses of the Cognitive Abilities Test (a different outcome measure not explicitly focused on attainment) found a smaller positive impact. Moreover, in terms of this outcome it appears as though disadvantaged pupils reaped fewer benefits from P4C than other pupils. It is unclear from the evaluation why there are these differences between the two outcomes.
4. Teachers reported that the overall success of the intervention depended on incorporating P4C into the timetable on a regular basis. Otherwise there was a risk that the programme would be crowded out.
5. Teachers and pupils generally reported that P4C had a positive influence on wider outcomes such as pupils' confidence to speak, listening skills, and self-esteem. These and other broader outcomes are the focus of a separate evaluation by the University of Durham.

Limitations

The evaluation is a reasonably large-scale trial in terms of number of schools and pupils involved. The time scale adopted was one complete calendar year which was quite a substantial amount of time that allowed the intervention to develop fully. However, this may still be too short a period for the kind of impact sought by the developers. The evaluation results have a limitation stemming from the design in that schools, rather than pupils, are randomised – reducing the 'power' of the study. There is no school dropout. The KS1 and matched KS2 results are from the National Pupil Database and include all cases for which there are records. Although attrition was kept to a minimum for the pre- and post-test CAT scores and all cases of missing data were pursued vigorously, the fact that 10% of individual pupil scores in the two year cohorts cannot be matched in the pre- and post-test CAT scores means that these specific results must be interpreted with some caution.

Interpretation

The objective of this randomised controlled trial was to test the gains in the academic performance of pupils, and also their cognitive ability, after taking part in P4C for one complete academic year. The trial was reasonably large, well conducted, retaining the complete school sample and 90% of the pupil sample from both year cohorts, and covering the process evaluation of the intervention and testing. This suggests that the results are reasonably secure.

It is clear that P4C, whatever its other possible benefits in terms of wider outcomes, does not hinder children's attainment at KS2 or their progress on CAT scores. In fact, in maths and reading there is a discernible but small benefit at KS2, equivalent to about two months of extra progress. All other indicators are positive but smaller (with the score for KS2 writing close to zero). Teachers and pupils generally report improved behaviour and relationships. This is achieved at a cost of around £16 per pupil. If there are wider or longer-term benefits to studying philosophy at primary school then this could make the intervention cost-effective. However, we do not yet know about these benefits. And there may be difficulties for some schools in adapting the existing setup to the demands of the intervention, especially if attempted as a whole-school process.

Analysis of the subgroups suggests that P4C is more effective with FSM-eligible pupils in terms of attainment, and for non-FSM and higher-attaining pupils at the outset in terms of CAT scores. The former suggests that P4C can be used to reduce the attainment gap in terms of poverty in the short term, but more investigation is needed to explain why the results for CAT scores are different.

Future research and publications

This project is a large-scale randomised controlled trial with no attrition at the school level. The evidence achieved from this project is an important contribution to the existing research on P4C and similar educational interventions. However, during the process of research we have developed the following research questions that still need to be answered.

P4C and the wider outcomes of education

During the process evaluation both pupils and teachers frequently mentioned their changing relationship with each other and with their peers. The questions developed on the basis of these observations are:

- Does P4C have an impact on pupil–teacher, pupil–pupil, and teacher–teacher relationships?
- Does P4C have an impact on pupils' confidence, well-being, and self-esteem?
- Does P4C have an impact on pupil 'voice' and levels of engagement with different/contrasting opinions – outside P4C settings and longer term?
- Does P4C have an impact on disadvantaged pupils' inclusion and class engagement?

The evaluation team at Durham University has developed a good relationship with SAPERE (the organisation for P4C) and the schools involved in the trial. We are utilising this working relationship between the developers and schools by conducting a follow-up in-depth study of P4C impact on wider outcomes. This research is sponsored by the Nuffield Foundation and over the next two years will continue to research non-cognitive impacts in the waitlist control schools who began P4C from September 2014.

Teachers' training and impact

How is the impact of P4C mediated by the way in which teachers run the sessions? P4C is largely an open approach and can be adopted by anyone. Is the training necessary? How much difference do specific P4C resources make? Such research questions could be answered through future studies.

P4C progress indicators

What instruments can best capture pupils' progress in P4C? The intervention deals with exploring concepts, understanding judgement criteria, evaluating problems that are based in a certain context, and developing the skills for collaborative argumentation. Key stage results and even CAT scores are not necessarily the most sensitive way to assess these skills. There is a need to find or develop an instrument that can measure pupils' progress in skills related to developing concepts and building arguments through reasoning. Two potential starting points for such an instrument may be the Cornell Test of Critical Reasoning and the Torrance Test of Creativity.

Future publications

The evaluators will prepare an article based on these results for a peer-reviewed journal.

References

Adey, P. and Shayer, M. (1994) *Really raising standards: Cognitive intervention and academic achievement*, London: Routledge

Berk, R. and Freedman, D. (2001) 'Statistical assumptions as empirical commitments', http://www.stat.berkeley.edu/~census/berk2.pdf (accessed 3 July 2014)

Campbell, J. (2002) *An evaluation of a pilot intervention involving teaching philosophy to upper primary children in two primary schools, using the Philosophy for Children methodology*, PhD thesis: University of Dundee

Carver, R. (1978) 'The case against statistical significance testing', *Harvard Educational Review*, 48, 378–399

Colom, R., Moriyón, F., Magro, C. and Morilla, E. (2014) 'The long-term impact of philosophy for children: A longitudinal study (preliminary results)', *Analytic Teaching and Philosophical Praxis*, 35, 1

Doherr, E. (2000) *The demonstration of cognitive abilities central to cognitive behavioural therapy in young people: Examining the influence of age and teaching method on degree of ability*, PhD thesis: University of East Anglia

Domitrovich, C.E. and Greenberg, M.T. (2000) 'The study of implementation: Current findings from effective programs that prevent mental disorders in school-aged children', *Journal of Educational and Psychological Consultation*, 11, 2, 193–221

Falk, R. and Greenbaum, C. (1995) 'Significance tests die hard: The amazing persistence of a probabilistic misconception', *Theory and Psychology*, 5, 75–98

GL Assessment Website: http://www.gl-assessment.co.uk/products/cat4-cognitive-abilities-test-fourth-edition

Gorard, S. (2013) *Research design*, London: Sage

Gorard, S. (2015) 'Rethinking "quantitative" methods and the development of new researchers', *Review of Education* (forthcoming)

Institute for the Advancement of Philosophy for Children (2002) 'IAPC research: experimentation and qualitative information, in Trickey, S. and Topping, K.J. (2004) 'Philosophy for children: A systematic review', *Research Papers in Education*, 19, 3, 365–380

Lipman, M. (1976) 'Philosophy for children', *Metaphilosophy*, 7, 1, 17–33

Lipman, M. (1990) 'Socrates for 6 year olds', BBC documentary (retrieved from https://www.youtube.com/watch?v=fp5lB3YVnlE)

Lipman, M., Sharp, A. and Oscanyon, F. (1980) Appendix B. *Philosophy in the classroom*, Philadelphia: Temple University Press

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*, Washington DC: Institute of Education Sciences

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

Stein, A., Haynes, J. and Unterstein (2003) *Assessing critical thinking skills*, Contribution to SACS/COC Annual Meeting, Nashville, Tennessee

Topping, K. J. and Trickey, S. (2007) 'Collaborative philosophical inquiry for schoolchildren: Cognitive gains at 2-year follow-up', *British Journal of Educational Psychology*, 77(4), 787–796

Trickey, S. and Topping, K. (2004) 'Philosophy for children: A systematic review', *Research Papers in Education*, 19, 3, 365–380

Watts, D. (1991) 'Why is introductory statistics difficult to learn?' *American Statistician*, 45, 4, 290–291

Williams, S. (1993) *Evaluating the effects of philosophical enquiry in a secondary school*, The Village Community School Philosophy for Children Project

Appendix 1: Research contract

P4C Research Project Contract

Please sign both copies, retaining one and returning the second copy to Jan Marples at SAPERE, Culham Innovation Centre, D5 Culham Science Park, Abingdon, Oxfordshire, OX14 3DB

Agreement to participate in the Evaluation of Philosophy for Children Research Project

School Name: _____

Aims of the Evaluation

The aim of this project is to evaluate the impact of the Philosophy for Children Research Project, a investigation of the effects of doing Philosophy for Children (P4C) on children's scores in SATs and Cognitive Abilities Tests. The results of the research will contribute to our understanding of what works in raising the pupil's attainment and will be widely disseminated to schools in England. Ultimately we hope that the evaluation will equip school staff with the strategies, skills and resources to use P4C develop children's reasoning abilities, their curiosity and positive attitudes towards school, learning and conversing with others.

The Project

The EEF will commission a team of researchers to test the effects of the doing Philosophy for Children on the achievement of over 5,000 KS2 pupils. There will be 5 groups of schools, each comprising 10 schools. Once the groups are established, EEF will assign all schools, randomly, to one of two categories: first phase schools or second phase schools. First phase schools will begin training in January 2013 and continue doing P4C for two years. Research will be carried out in both sets of schools using pre- and post-intervention tests. Test results will be compared and differences noted. Second phase schools will begin training in September 2014 and continue with P4C after that. The random selection of schools ensures that the research is a fairer test of P4C than it would otherwise be. The continuing involvement of all schools is critical to the project s success so, in order to become part of the project, it is necessary to accept the randomisation process.

Structure of the Evaluation

The evaluation is being conducted by Birmingham University. Schools that agree to take part are randomly allocated to either a first phase of second phase (control) group.

Schools in the first phase group receive training and support beginning in January 2013.

Schools in the second phase group receive training and support beginning in September 2014.

In the evaluation scores for SAT and Cognitive Abilities tests will be compared for all pupils. In addition other data will be collected using surveys and interviews.

For each pupil in years 4 to 6 in all participating schools, we will need access to the following:

1. Their unique pupil id, as used in the National Pupil Database (NPD), in order to link data on each pupil at an individual level. The link between the pupil id and any identifying information will be held separately from all other data, and will be destroyed immediately after use. The project does not require to know, and does not wish to know, who each pupil is. But it does need to be able to match data anonymously at an individual pupil level.
2. Their eventual Key Stage 2 results (levels and points) for each subject.
3. Their NPD record, especially their KS1 results (levels and points), gender, month of birth, FSM status, SEN status, ethnicity, and first language.
4. Their attendance record, date of leaving (if during the project), and any exclusions (where applicable).

For each pupil in years 4 and 5 in all participating schools, we will also need access to the following: CAT4 scores in December 2012 and December 2013. The CAT4 will be administered during a convenient lesson time as part of the project.

RESPONSIBILITIES

SAPERE WILL:

- DELIVER TRAINING SESSIONS, SUPPORT AND RESOURCES

  First phase schools. Two initial days of training (or equivalent) plus six support days in year one. Three support days in year two plus four days of training for two teachers. There will also be support via access to recourses and online advice. Supply costs for the training in year two will be covered by the project assuming at a rate of £175 per teacher.

  Second phase schools. Two initial days of training (or equivalent) plus three support days in year one. One support days in year two plus four days of training for two teachers. There will also be support via access to recourses and online advice. Supply costs for the training in year two will be covered by the project assuming at a rate of £175 per teacher.

- BE THE FIRST POINT OF CONTACT FOR ANY QUESTIONS ABOUT THE EVALUATION
- SEND OUT REGULAR UPDATES ON THE PROGRESS OF THE PROJECT THROUGH A NEWSLETTER
- ENSURE ALL TRAINERS HAVE RECEIVED CRB CLEARANCE

THE EVALUATION TEAM WILL:

- COLLECT AND ANALYSE ALL THE DATA FROM THE PROJECT
- ENSURE ALL STAFF CARRYING OUT ASSESSMENTS ARE TRAINED AND HAVE RECEIVED CRB CLEARANCE
- PROVIDE HEAD TEACHERS WITH ALL ATTAINMENT DATA AFTER THE TESTS HAVE BEEN COMPLETED
- DISSEMINATE RESEARCH FINDINGS

THE SCHOOL WILL:

- Allow time for each testing phase and liaise with the evaluation team to find appropriate dates and times for testing to take place
- In year one organise 2 days of training in a combination of closure days and after school sessions

- Release 2 KS2 Staff so that they can attend training sessions in year 2.
- Ensure the shared understanding and support of all school staff for to the project and personnel involved.
- Facilitate at least one hour per week of P4C for all children at Key Stage 2.
- Be a point of contact for parents / carers seeking more information on the project.

We commit to the Evaluation of Philosophy for Children Research Project as detailed above

Signatures

ON BEHALF OF THE SAPERE

PROJECT LEADER STEVE WILLIAMS: _____

DATE: _____

ON BEHALF OF THE EVALUATION TEAM:

LEAD EVALUATOR STEPHEN GORARD: _____

DATE: _____

ON BEHALF OF THE SCHOOL:

HEAD TEACHER (NAME AND SIGNATURE) _____

OTHER RELEVANT STAFF (NAME AND SIGNATURE): _____

DATE: _____

Appendix 2: Sub-group analyses of potential interest, not pre-specified in the protocol

Table 22: CAT4 gain score for those with higher CAT scores at the outset (≥ 94.8 in pre-test)

	N	Gain score	Standard deviation	'Effect' size
P4C	633	1.38	7.38	+0.14
Control	727	0.40	6.91	-
Total	1360	0.86	7.14	-

Table 23: CAT4 gain score for those with the lower CAT scores at the outset (< 94.8 in pre-test)

	N	Gain score	Standard deviation	'Effect' size
P4C	733	2.85	7.71	-0.02
Control	728	2.99	7.50	-
Total	1461	2.92	7.60	-

Table 24: Boys CAT4 gain score

	N	Pre-CAT4	Standard deviation	Post-CAT4	Standard deviation	Gain score	Standard deviation	'Effect' size
P4C	696	93.97	11.38	96.21	12.56	2.25	7.76	+0.07
Control	756	94.73	11.43	96.43	12.19	1.68	7.47	-
Total	1452	94.37	11.41	96.33	12.37	1.96	7.61	-

Note: the 'effect' size for girls was +0.06

Table 25: SEN pupils CAT4 gain score

	N	Pre-CAT4	Standard deviation	Post-CAT4	Standard deviation	Gain score	Standard deviation	'Effect' size
P4C	231	86.75	8.65	87.57	9.54	0.82	6.72	+0.02
Control	284	88.71	9.23	89.38	10.73	0.67	7.12	-
Total	515	87.83	9.02	88.57	10.24	0.74	6.94	-

Note: the 'effect' size for non-SEN pupils was +0.07

Table 26: Ethnic minority pupils CAT4 gain scores

	N	Pre-CAT4	Standard deviation	Post-CAT4	Standard deviation	Gain score	Standard deviation	'Effect' size
P4C	417	94.55	10.60	97.57	11.95	3.03	7.18	+0.13
Control	342	94.93	12.20	96.99	12.95	2.06	7.79	-
Total	759	94.72	11.34	97.31	12.40	2.59	7.47	-

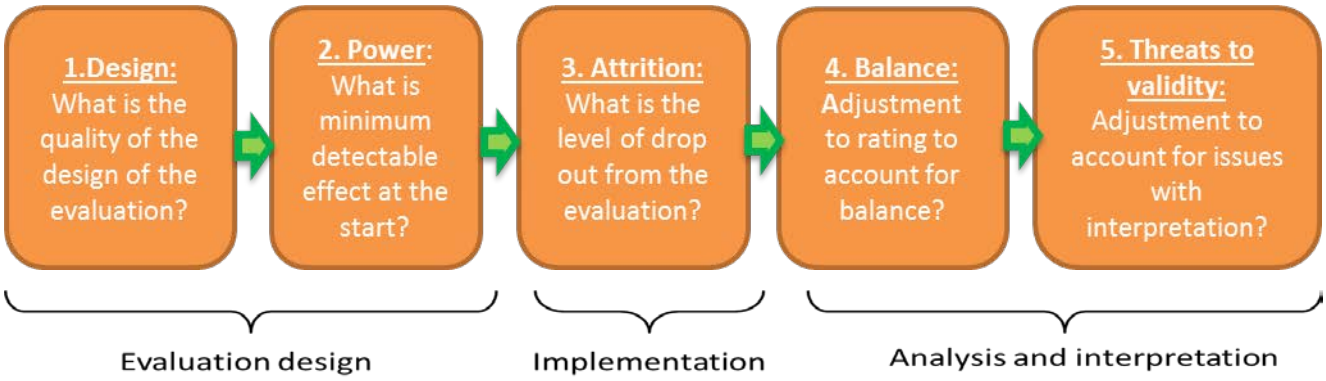
Note: the 'effect' size for White UK pupils was +0.02

Table 27: Pupils with EAL CAT4 gain scores

	N	Pre-CAT4	Standard deviation	Post-CAT4	Standard deviation	Gain score	Standard deviation	'Effect' size
P4C	134	95.52	12.04	99.19	12.89	3.66	6.99	+0.19
Control	244	94.01	11.76	96.29	12.60	2.28	7.50	-
Total	378	94.55	11.87	97.32	12.76	2.77	7.35	-

Note: the 'effect' size for non-EAL pupils was +0.05

Appendix 3: Security classification of trial findings



Rating	1. Design	2. Power (MDES)	3. Attrition	4. Balance	5. Threats to validity
5 🔒	Fair and clear experimental design (RCT)	< 0.2	< 10%	Well-balanced on observables	No threats to validity
4 🔒	Fair and clear experimental design (RCT, RDD)	< 0.3	< 20%		
3 🔒	Well-matched comparison (quasi-experiment)	< 0.4	< 30%		
2 🔒	Matched comparison (quasi-experiment)	< 0.5	< 40%		
1 🔒	Comparison group with poor or no matching	< 0.6	< 50%		
0 🔒	No comparator	> 0.6	> 50%	Imbalanced on observables	Significant threats

The final security rating for this trial is 3 🔒. This means that the conclusions have moderate security.

This evaluation was designed as a randomised controlled trail. The sample size was designed to detect a MDES of less than 0.4, by design, reducing the security rating to 3 🔒. At the unit of randomisation (school), there was zero attrition, and extremely low attrition at the pupil level also. The post-tests were administered by the schools by teachers who were aware of the treatment allocation, but with invigilation from the independent evaluators. Balance at baseline was high, and there were no substantial threats to validity.

Appendix 4: Cost rating

Cost rating	Description
£	*Very low:* less than £80 per pupil per year.
£ £	*Low:* up to about £200 per pupil per year.
£ £ £	*Moderate:* up to about £700 per pupil per year.
£ £ £ £	*High:* up to £1,200 per pupil per year.
£ £ £ £ £	*Very high:* over £1,200 per pupil per year.

This document is available for download at **www.educationendowmentfoundation.org.uk**

Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk