Journal of Philosophy of Education
The Journal of Philosophy of Education Society of Great Britain
pesgb

Journal of Philosophy of Education, Vol. 51, No. 1, 2017

Can 'Philosophy for Children' Improve Primary School Attainment?

STEPHEN GORARD, NADIA SIDDIQUI AND BENG HUAT SEE

There are tensions within formal education between imparting knowledge and the development of skills for handling that knowledge. In the primary school sector, the latter can also be squeezed out of the curriculum by a focus on basic skills such as literacy and numeracy. What happens when an explicit attempt is made to develop young children's reasoning—both in terms of their apparent cognitive abilities and their basic skills? This paper reports an independent evaluation of an in-class intervention called 'Philosophy for Children' (P4C), after just over one year of schooling. The intervention aims to help children become more willing and able to question, reason, construct arguments and collaborate with others. A group of 48 volunteer schools were randomised to receive P4C (22 schools) or act as a control for one year (26). This paper reports the CAT results for all pupils in years 4 and 5 initially, and the Key Stage 2 attainment in English and Maths for those starting in year 5. There was no school dropout. Individual attrition from a total of 3,159 pupils was around 11 percent—roughly equal between groups. There were small positive 'effect' sizes in favour of the P4C group in progress in reading (+0.12) and maths (+0.10), and even smaller perhaps negligible improvements in CAT scores (+0.07) and writing (+0.03). The results for the most disadvantaged (free school eligible) pupils were larger for attainment (+0.29 in reading, +0.17 writing and +0.20 maths), but not for CATs (–0.02). Observations and interviews suggest that the intervention was generally enjoyable and thought to be beneficial for pupil confidence. Our conclusion is that, for those wishing to improve attainment outcomes in the short term, an emphasis on developing reasoning is promising, especially for the poorest students, but perhaps not the most effective way forward. However, for those who value reasoning for its own sake, this evaluation demonstrates that using curriculum time in this way does not damage attainment (and may well enhance it and reduce the poverty gradient in attainment), and

Published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

so suggests that something like P4C is an appropriate educational approach.

BACKGROUND

Philosophy for Children (P4C) was developed from an initiative by Professor Matthew Lipman in New Jersey, USA in 1970 with the establishment of the Institute for the Advancement of Philosophy for Children (IAPC). P4C has since become a worldwide educational approach, and something like it has been adopted by schools in 60 countries across the world, although the nature of the practice varies (Mercer *et al.,* 1999). In the UK, the Society for the Advancement of Philosophical Enquiry and Reflection in Education (SAPERE) was established in 1992. SAPERE promotes the use of P4C in UK schools along with developing teaching resources and providing teacher training courses. Lipman's central idea of creating a classroom 'community of enquiry' is retained along with the wider sequence of activities and materials that constitute a P4C session. However, many of the materials used are original to the SAPERE version.

An initial evaluation of the original Philosophy for Children scheme was conducted by 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. A matched comparison group design was used in this evaluation, in which 20 pupils received the intervention and their counterpart group of 20 students were taught social studies in a traditional way in the same amount of time. The report does not explain how matching was done. The study reported significant gains in logical reasoning and reading, measured using the California Test of Mental Maturity (CTMM). Differences in reading scores were reported to have been maintained 2.5 years later.

A second, larger experiment reported in Trickey and Topping (2004) involved 200 pupils. Sessions were conducted 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. The process of school selection and allocation in treatment and control groups were not clearly specified.

A systematic review was conducted by Trickey and Topping (2004) which showed consistent moderate effects on a range of outcome measures. The mean effect size for the studies included was 0.43. However, these studies were not always fully comparable because of the different outcomes measured and the different instruments used for measuring them. For example, 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 more generic tests of literacy and numeracy.

One of the earliest studies in the UK was conducted by Williams (1993). The study examined the effects of 27 one-hour P4C lessons (using Lipman's materials) on reading comprehension, reasoning skills and intellectual confidence. Participants were 42 pupils from two Year 7 classes in one school in Derbyshire, UK. Results were obtained for 32 children. Children 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 the P4C group made significantly bigger gains than control pupils. Significant gains were also reported for reasoning skills and intellectual confidence. These were measured using bespoke evaluation tools and video recordings of pupils' interaction during lessons which the evaluators had to make subjective judgements about. Nevertheless, 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 TRAC programme (Talk, Reasoning and Computers) 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 10 weeks. The study involved 60 Year 4 and 5 pupils (age 9 to 10 years) 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. Experimental pupils made significantly bigger gains between pre- and post-test compared to control pupils.

A recent randomised controlled trial was conducted with 540 pupils in years 7 and 8 (Fair *et al.*, 2015). The study found high positive gains in CAT scores for year 7 pupils who received the treatment against their equivalent controlled pupils who were taking language arts classes. The equivalent gains for year 8 pupils were much lower. This could be because pupils in year 7 were exposed for 26 weeks and pupils in year 8 were given P4C session for only 10 weeks. The difference in gains between the two groups was attributed to the difference in dosage. The study does not clarify the baseline equivalence assessment of pupils who were allocated in two groups and there is an indication that pupils were not equally balanced as pupils in the treatment group were ahead of their counterparts in CAT pre-test scores.

Despite this evidence of short-term improvements, some commentators have 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 were available for only 115 pupils. Experimental pupils received 1 hour per week of collaborative enquiry lessons, while control pupils continued regular lessons. After 16 months of intervention (with 1 hour of P4C per week) the treatment pupils made substantial improvements in test scores whereas control pupils performed worse than when they started at pre-post-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 (Table 3, p. 794).

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 from two private schools over 20 years. 455 children aged 6 years (first year of primary school) to 18 years (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 when children were aged 8, 11/12 and 16 years. 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 to lower ability pupils, but this was not clear from their presentation of the analysis. Moreover, although large scale and long term, the students were not randomised in terms of receiving P4C instruction, and the study may not be generalisable as pupils came from relatively prosperous families in private schools. 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 (as in Tok and Mazi, 2015), 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 if philosophical enquiry can lead to enhanced performance in academic domains and whether it would have the same impact in UK schools with British children. 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.

The main aim of the current 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. 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.

P4C PRACTICE IN CLASSROOMS

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. Pupils and teacher sit in a circle so everyone can see and hear one another. The teacher negotiates with pupils on guidelines on the conduct of sessions and the purpose is to set some basic rules of communication agreed all the pupils.

The teacher then introduces the planned material she/he has chosen in order to provoke pupils' interest, puzzle them or prompt their sense of what is important. A minute of silence is 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.

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. When the listing of questions is complete, the next phase is to select one as a dialogue starter. The selection is made by pupils using one of a range of voting methods. The discussion floor is then open for all to share their views.

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. Teachers will often prompt pupils to imagine alternatives and consequences, seek evidence, quantify with expressions like 'all', 'some' or 'most', offer examples and counter examples and question assumptions.

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 practice arguments with peers before sharing with the whole group. Teacher can also draw diagrams or make notes to keep track of significant arguments.

The closing of the session involves last words from all pupils. Pupils might have the same opinion as in the beginning or it could have changed as a result of 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 a detailed reflection whereby a teacher could ask pupils to write a summary of their views.

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 on in the next P4C sessions.

P4C, as promoted by SAPERE, is a template to practice and organise a classroom session for philosophical enquiry. It does not have any specified materials or stimuli that must be used; there are only examples and suggestions. The steps 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.

No special equipment is required for this intervention. It may involve standard material for teaching such as a projector, board, pens and paper. All pupils and teachers are required to sit comfortably in a circle facing each other for discussion. There is also the expectation that teachers will use existing curriculum material in their lessons when they judge it to have the potential to stimulate philosophical discussion and clarify key concepts in subject areas such as democracy, justice, nation, history, truth, cause, evidence, beauty, art, real, belief, knowledge, tolerance and theory.

Control Group Activity

The control schools (on the waiting list) were funded and permitted to receive P4C teacher training and implement the intervention after the trial was 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 'clean' control situation but the evaluators visited control schools and were not aware of any systematic approach to critical thinking adopted in the control schools.

Cost

The P4C foundation course is the initial training for teachers and teaching assistants after which they can implement the intervention at a whole school level. For the current evaluation, these costs were met by Educational Endowment Foundation (EEF) both for the treatment schools and later for the waiting-list control schools. SAPERE states that its current programmes typically cost £25–30 per pupil.

Methods of Investigation

The study is a randomised controlled trial with schools allocated to one of two arms receiving the P4C intervention over one year or not. Pupils involved in the intervention were in Key Stage 2 (Years 3 to 6), and all formed part of the process evaluation. 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.

A total of 48 primary schools were recruited from London, Hull, Sheffield, Manchester, Hertfordshire, Staffordshire and Stoke-on-Trent in England. None had prior experience of using P4C. All schools had at least 25 percent of their pupils known to be eligible for free school meals. At least 10 of the schools had fewer than 60 percent of pupils achieving Level 4+ in English and maths, and with pupils making below-average progress in English and maths in 2011.

Table 1. Percentage of pupils with specified background characteristics in each group

	Intervention	Control
Male	51	52
FSM-eligible	48	46
SEN reported	18	19
English as additional language	9	15
Non-white UK ethnicity	31	23

It was planned that 22 of these schools would be in the intervention group from September 2012, and they were randomised at school level accordingly. 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 control schools were funded to receive P4C from September 2014. There was no school dropout. The two groups were well-balanced in terms of sex, their eligibility for free school meals (FSM—a measure of family poverty) and their special needs status (SEN—an indicator of a specific learning difficulty or challenge) (Table 1).

Opt-out consent forms were then sent by schools 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. A total of 3,159 pupils in Years 4 and 5 (entire year groups) were in schools taking part in the trial at the outset, of which 1,550 were in the treatment group and 1,609 in the control group. Traditional power calculations to estimate the minimum sample size required make a number of assumptions that are not relevant here (such as no dropout of cases after randomisation to treatment or control), 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 group (treatment or control) would be needed (for 80 percent power to detect a difference of 0.4 with alpha of 5 percent) according to Lehr's formula (Gorard, 2013a). In fact, the situation is better than this, because of the correlation between pre- and post-test scores for Key Stage 1 and Key Stage 2 data, and for CAT4 (see below). Thus, a sample of 48 schools with over 3,000 pupils should easily provide sufficient traditional 'power' to detect an effect in terms of either outcome. Around 11 percent of pupils with pre-test scores are missing a post-test score for CAT4 (see Figure 1).

The main outcomes of interest in assessing the impact were the English and maths 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, and the Cognitive Abilities Test (CAT4) scores of all initial year 4 and 5 pupils.

The individual results for Key Stage 2 reading, writing and maths 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. Because the Key Stage 1 pre-scores and Key Stage 2 post-scores were on different metrics both were converted to z-scores to assist comparability.

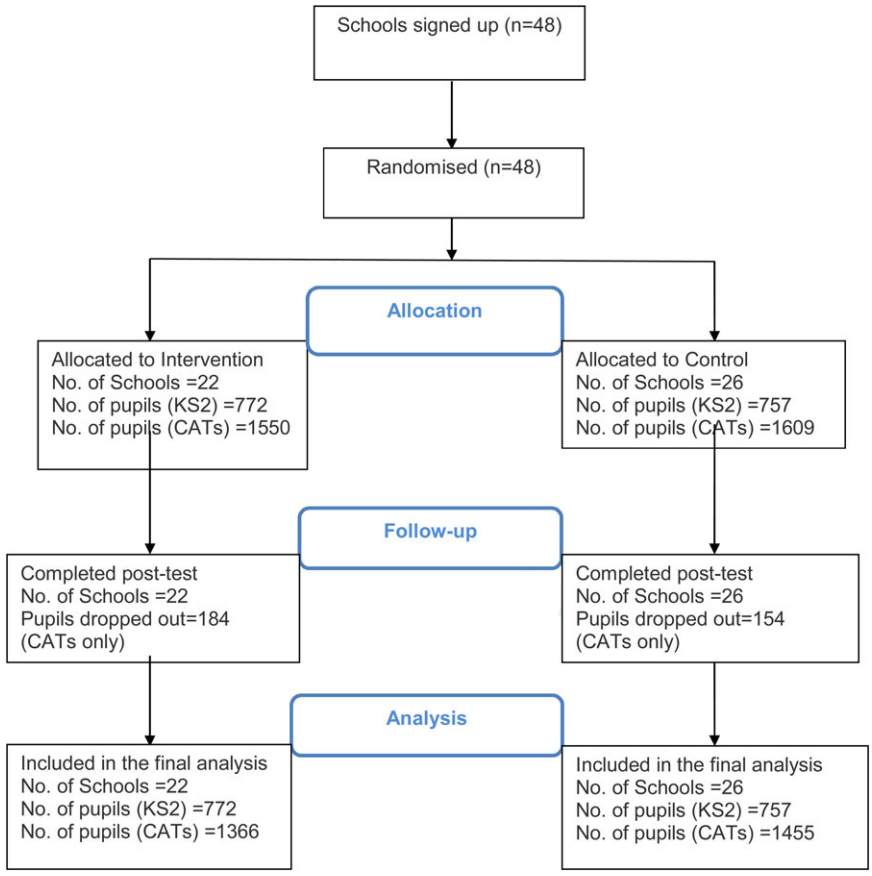

Figure 1: Flowchart of number of cases at start and end of trial

The online CAT4 test was proposed by SAPERE as it was judged appropriate for the kinds of things that P4C might influence, and it had previously been used in the Topping and Trickey (2007) study. The test has four sub-scales, representing the core elements thought to be needed for critical thinking (Stein *et al.*, 2013). These are verbal, non-verbal, quantitative and spatial ability. If P4C has an impact it is more likely to be on verbal than spatial skills, for example.

The effectiveness or otherwise of P4C is represented by:

- the effect size (Hedges' *g*) for the standardised gain score from Key Stage 1 to Key Stage 2 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 sub-scales—verbal, quantitative, non-verbal and spatial.

Additional analyses were performed, repeating the overall analysis but using scores for only those pupils eligible for FSM (as pre-specified), for each year group, and those above or below the median (middle) score in the CAT pre-test (GL Assessment, 2016).

The evaluation team made 30 trips 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 an informal conversation 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 IMPACT RESULTS

Headline Results

At the outset the treatment and control groups were reasonably balanced, with the control group having slightly better Key Stage 1 scores in each of reading, writing and maths (Tables 2 to 4). By the end the treatment group had narrowed this gap in all three subjects, especially reading and maths

Table 2. Key Stage 1 to Key Stage 2 reading progress, by group

	N	Mean Key Stage 1 points z-score	SD	Mean Key Stage 2 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	1,529	0	1	0	1	0	0.90	–

Table 3. Key Stage 1 to Key Stage 2 writing progress, by group

	N	Mean Key Stage 1 points z-score	SD	Mean Key Stage 2 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	1,529	0	1	0	1	0	0.84	–

Table 4. Key Stage 1 to Key Stage 2 maths progress, by group

	N	Mean Key Stage 1 points z-score	SD	Mean Key Stage 2 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	1,529	0	1	0	1	0	0.78	–

Key Stage 2 scores. For this reason, these results are all presented as gain scores, representing progress from Key Stage 1 to Key Stage 2.

Viewed as gain scores there is evidence here that P4C might have a positive impact on pupil attainment at Key Stage 2, 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. The practice of P4C involves reading and oral reasoning. The results in Tables 2 and 4 are unlikely to be due to chance, or bias due to missing data. The number of counterfactual cases that would need to be added to the smaller (control group) in order for the effect sizes to be zero would be 91 and 76 (see Gorard and Gorard, 2016).

The two groups were also reasonably well balanced in terms of CAT scores at the outset, but again with the control slightly ahead (Table 5). For this reason the results are presented as 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 months' extra progress in just over a year).

The number of counterfactual cases needed to disturb this finding would be 96, but unlike with the Key Stage 2 results from the National Pupil Database, this figure has to be compared with the number of pupils with

Table 5. Overall CAT4 gain score

	N	Pre-CAT4	Standard deviation	Post-CAT4	Standard deviation	Gain score	Standard deviation	'Effect' size
P4C	1,366	94.37	11.24	96.59	12.26	2.22	7.59	+0.07
Control	1,455	95.20	11.19	96.90	11.90	1.70	7.32	–
Total	2,821	94.80	11.22	96.75	12.07	1.95	7.46	–

Table 6. 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	1,360	0.86	7.14	–

Table 7. 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	1,461	2.92	7.60	–

missing post-test scores (see Figure 1). Clearly, P4C is doing no harm to pupils' attainment or cognitive attainment, and there is some promise here. However, the CAT 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. P4C showed the biggest 'impact', on average, in terms of the verbal sub-scale (+0.08). This is both to be expected and ties in with the greater gain for the treatment group in Key Stage 2 reading.

The use of post-test only scores is both simpler and to be preferred over comparing pre- and post-tests scores as here, for a number of good reasons (Gorard, 2013b). This was not possible here because of the slight imbalance in the two group scores at the outset. Although the differences were neither large nor extreme, as would be required for the process of 'regression to the mean' to explain the results, some readers may still be concerned about this. Tables 6 and 7 show that it was the pupils in the higher-scoring half of CAT scores at the outset (at or above the median score) who created the overall positive result. The lower scoring half made no apparent progress. This demonstrates that the headline result cannot be explained by regression to the mean.

Sub-group Analyses

Analyses of selected sub-groups of the randomised pupils, as here, do not have the force of a trial. However, they are useful in providing a possible

Table 8. Key Stage 1 to Key Stage 2 Reading progress—FSM-eligible pupils only

	N	Mean Key Stage 1 points z-score	SD	Mean Key Stage 2 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. Key Stage 1 to Key Stage 2 Writing progress—FSM-eligible pupils only

	N	Mean Key Stage 1 points z-score	SD	Mean Key Stage 2 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. Key Stage 1 to Key Stage 2 maths progress—FSM-eligible pupils only

	N	Mean Key Stage 1 points z-score	SD	Mean Key Stage 2 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	–

Table 11. No CAT4 gain score—FSM-eligible pupils only

	N	Gain score	Standard deviation	'Effect' size
P4C	697	1.45	7.17	–0.02
Control	781	1.66	7.36	–
Total	1,478	1.56	7.27	–

explanation and context for the headline results. Tables 8 to 10 show the results (pre-, post- and gain scores) for only those pupils known to be eligible for free school meals (FSM). The 'effect' sizes are more positive than for Tables 2 to 4, suggesting that P4C is effective for FSM-eligible pupils, and that P4C could be one way of reducing the current poverty gradient in Key Stage 2 results. The number of counterfactual cases needed to disturb each finding would be 68, 40 and 47, respectively (considerably higher than the number of FSM-eligible pupils with missing scores).

The same result does not appear for the CAT scores (Table 11). 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 Key Stage 2 attainment.

REPORTED OUTCOMES

Implementation

The implementation of P4C in the schools was closely monitored by SAPERE to ensure 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.

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 venues for conducting P4C sessions.

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 for 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:

- Is it acceptable for people to wear their religious symbols at work places?
- Are people's physical looks more important than their actions?
- What is kindness?
- Can you and should you stop free thought?
- Is it OK to deprive someone of their freedom?

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.

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 not seen to be as important a part of 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.

It was observed by the evaluators that P4C sessions should usually be a complete sequence of steps, 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 and sufficiently reviewed. The session was rushed to the end as the time for the session was passing quickly. As an observer it was felt that the 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 pupils might 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.

Teacher Feedback on P4C

According to the teachers who were asked about the challenges of implementing P4C, it was understood that the success of the intervention depended on incorporating P4C in 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. 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.

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 they had 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'.

The intervention is attractive to stakeholders such as 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 sub-groups as the need arises.

The teachers' reflection on the training informs that the training was essential because the intervention is based more on exploring concepts rather than just doing 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 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 is 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 informed that they felt more relaxed talking during P4C compared to normal lessons because they knew that teacher would not discourage them from talking and discussing.

It was observed by the evaluators in different P4C sessions that it helps pupils' confidence and engagement if teachers are equal participants in the enquiry circle rather than from their position of authority in the classroom. In some earlier 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 let pupils talk and discuss more. However, from observations made in the later period it was noticed that the same teachers moderated the sessions which were well balanced in terms of their own and as well as pupils' participation.

Pupil Feedback

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 get to know what their peers think during P4C, which is not so possible in any other school situation. 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 the pupils.

The most common thing 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 would also have preferred spending P4C time doing activities like sport, while a small number would have preferred a 'normal' lesson.

DISCUSSION

The reported evaluation was a large-scale trial in terms of the number of schools and pupils involved. The time scale adopted was over 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 Key Stage 1 and matched Key Stage 2 result are

from the National Pupil Database and include all cases for which there are records.

It is clear that P4C, whatever its other possible benefits in terms of wider outcomes, does not hinder children's attainment at Key Stage 2. In fact, in maths and reading there is a discernible but small benefit at Key Stage 2, perhaps equivalent to about two months of extra progress. All other indicators are positive but even smaller (with the score for Key Stage 2 writing close to zero). Teachers and pupils generally report improved behaviour and relationships. This is achieved at a cost of around £30 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 are some difficulties in adapting the existing school setup in some schools to the demands of the intervention, especially if attempted as a whole-school process.

Correspondence: Stephen Gorard, Nadia Siddiqui and Beng Huat See, School of Education, Durham University, Durham, DH1 1TT, UK.
Email: s.a.c.gorard@durham.ac.uk

REFERENCES

Adey, P. and Shayer, M. (1994) *Really Raising Standards: Cognitive Intervention and Academic Achievement* (London, Routledge).

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, mimeo, University of Dundee.

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, pp. 50–56.

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, mimeo, University of East Anglia.

Fair, F., Haas, L., Gardosik, C., Johnson, D., Price, D. and Leipnik, O. (2015) 'Socrates in the Schools from Scotland to Texas: Replicating a Study on the Effects of a Philosophy for Children Program'. *Journal of Philosophy in Schools*, 2.1, pp. 18–37.

GL Assessment Website (2016) Cognitive Abilities Test (Swindon, GL Assessment). Available online at: http://www.gl-assessment.co.uk/products/cat4-cognitive-abilities-test-fourth-edition.

Gorard, S. (2013a) *Research Design* (London, Sage).

Gorard, S. (2013b) The Propagation of Errors in Experimental Data Analysis: A Comparison of Pre- and Post-Test Designs. *International Journal of Research and Method in Education*, 36.4, pp. 372–385.

Gorard, S. and Gorard, J. (2016) What to Do Instead of Significance Testing? Calculating the Number of Counterfactual Cases Needed to Disturb a Finding. *International Journal of Social Research Methodology*, 19.4, pp. 481–490.

Institute for the Advancement of Philosophy for Children (2002) IAPC Research: Experimentation and Qualitative Information, in Trickey, S. and Topping, K.J. *Philosophy for Children: a Systematic Review, Research Papers in Education*, 19.3, pp. 365–380.

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

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, pp. 95–111.

Moriyón, F. and Tudela, E. (2004) What we Know about Research in Philosophy with Children. Available online at: https://philoenfant.org/2015/10/30/resume-de-103-recherches-en-philosophie-pour-les-enfants/. Last accessed: 25 June 2016.

Stein, A., Haynes, F. and Unterstein, J. (2003) *Assessing Critical Thinking Skills*, Contribution to SACS/COC Annual Meeting, Nashville, Tennessee.

Trickey, S. and Topping, K. (2004) Philosophy for Children: A Systematic Review, *Research Papers in Education*, 19.3, pp. 365–380.

Tok, Ş. and Mazı, A. (2015) The Effect of Stories for Thinking on Reading and Listening Comprehension: A Case Study in Turkey, *Research in Education*, 93.1, pp. 1–18.

Topping, K. and Trickey, S. (2007) Collaborative Philosophical Inquiry for Schoolchildren: Cognitive Gains at 2-year Follow-up, *British Journal of Educational Psychology*, 77.4, pp. 787–796.

Williams, S. (1993) *Evaluating the Effects of Philosophical Enquiry in a Secondary School* (The Village Community School Philosophy for Children Project). Available online at: http://www.thinkingscripts.co.uk/pdf/villagep4c.pdf.