The Routledge International Handbook of Research on Teaching Thinking

Rupert Wegerif, Li Li, James C. Kaufman

Philosophy for children

Publication details
Keith J. Topping, Steve Trickey
Published online on: 03 Jun 2015


PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Increasing interest in the promotion of critical and creative thinking stems from discourse about the changing nature of the skills needed in contemporary society. In a number of countries, curricular prescriptions have shown decreased emphasis on content knowledge and increased emphasis on transferable skills such as critical and creative thinking and collaborative problem solving, both as a means to facilitate deeper learning and as important life skills. In parallel, such skills have been seen as part of the changing skill set required by employers.

Programmes to develop thinking skills can be divided into: separate programmes, subject-based programmes, infused programmes and multi-method programmes. Separate programmes are those which often take up school time but are not connected to the school curriculum. Subject-based programmes are those which are embedded within a particular school subject, such as science, mathematics or literacy. ‘Infused’ programmes are those which aim to infiltrate the whole of the school’s pedagogical activities. Multi-method programmes are those which could be placed in more than one of the other three categories.

The Philosophy for Children programme (Lipman & Sharp, 1978; Lipman, Sharp & Oscanyon, 1980; Lipman, 1981, 2003) has often been thought of as a separate programme. However, the methodology of Philosophy for Children (P4C) potentially can be infused into a range of subject domains, even though it initially tends to be used with children as a separate and novel activity. Consequently we would categorise it as a ‘multi-method’ programme.

What is ‘Philosophy for Children’?

Lipman’s Philosophy for Children (P4C) aims to teach children to think for themselves and make informed choices. The goal is to ‘improve children’s reasoning abilities and judgement by having them thinking about thinking as they discuss concepts of importance to them’ (Lipman, 1981, p.37). Lipman argued that the skills necessary for critical and creative thinking were acquired within the context of language (although this raises issues with children who are linguistically disadvantaged).

P4C is delivered through a set of core materials consisting of seven novels with accompanying manuals. It is introduced at 6–7 years of age, the assumption being that children are capable
of critical and reflective thinking even at this stage. The programme is delivered over a long
time span – from the age of 6 to 16 years. The novels serve as springboards for debate. Their
central characters learn to resolve their problems through their powers of reasoning. The class
acts as a ‘community of inquiry’. The skills of the teacher in providing open-ended questions in
a supportive emotional climate are crucial to the outcomes of the process.

Alternative stimulus materials have evolved since Lipman’s original materials. In the UK,
of materials used with the Clackmannanshire Thinking Skills initiative. In general, P4C involves
pupils and their teacher sharing a short story, picture, poem, object or some other stimulus. The
children then take time to think of their own questions. These are then discussed briefly before
one is selected for more extensive discussion.

The inquiry process should include teacher behaviours that Cotton (2002) suggested help to
foster a climate conducive to the development of thinking skills, including the setting of ground
rules well in advance, showing respect for each pupil, providing non-threatening activities,
accepting individual differences, modelling thinking skills, and allowing students to be active
participants. As Haynes (2002) pointed out, philosophical inquiry is not a ‘tool-kit’ approach to
promoting independent thinking. The process is dependent on the quality of interaction and
dialogue engendered, rather than rigidly following a step-by-step procedure.

Issues in the evaluation of Philosophy for Children

What constitutes an appropriate means of evaluating the outcomes and process of Philosophy
for Children has been the subject of debate. These issues potentially involve both formative and
summative aspects of assessment. Naturally, the goal of any process needs to be clarified before
it can be evaluated. Lipman (1980, p.15) proposed that an important goal of the Philosophy
for Children programme was that of helping children become more reasonable individuals so
that they were able to make wiser judgements. He suggested that the Philosophy for Children
approach encourages children to develop both critical reasoning and creative thinking skills.
This implies that progress in reasoning is an important element to evaluate.

There is much debate about whether an outcomes approach to evaluating the use of Philosophy
for Children is appropriate and if so what form such an evaluation should take. Lipman (1980, p.44)
recognised that there was a need to evaluate the outcomes of engaging children in philosophical inquiry
if teachers were to invest precious time and energy in regular philosophical inquiry. He wrote:

Some people have argued that humanistic studies should not be compelled to justify
themselves by virtue of empirical evidence that they promote academic achievement. But
the argument is not likely to be persuasive to school administrators who have to justify the
actual decisions as to which courses to introduce.

Philosophy would only be included in the curriculum ‘if it can demonstrate . . . that it can make
a significant difference to the overall performance’.

However, Reznitskaya (2005) has suggested that outcome studies of Philosophy for Children
are limited in that they ‘form an overall opinion regarding the success of a program, rather
than trying to understand the underlying mechanisms of cognition’ and are ‘ineffective for
advancing our understanding’ and ‘providing P4C practitioners with specific, empirically tested
instructional strategies’. This concern seems to reflect the different functions of formative and
summative assessments. Similarly, Reznitskaya (2005) voiced concern about the limitations of
standardised tests in evaluating individual gains in reasoning.

Reznitskaya’s concern seems to resonate with previous comments by Burden and Nichols
(2000) about the use of pre-post designs in evaluating cognitive programmes in general. Such
designs can be seen as an unrealistic attempt to replicate controllable conditions in complex real life situations such as classrooms – inevitably simultaneously influenced by large numbers of confounding variables. As a result, Burden and Nichols argue that ‘traditional pre-post designs are not the most appropriate means of evaluating the impact of curriculum interventions like cognitive education programs’ (p. 294). The analysis of transcripts of discussion taking place during enquiries offers an alternative approach to pre-post evaluations of Philosophy for Children. Baron (in Baron and Sternberg, 1987) discussed the relative merits and disadvantages of analysis of verbal output in discussion and written output.

Salmon (2003) considered the qualitative versus quantitative debate in terms of five features that he argued underpinned good quality research. While Salmon recognised that qualitative methods can claim comparative rigour, he suggested they tended to ‘duck the question’ of issues of generalisation and transferability and that they were ‘less helpful in inviting the reader to judge whether their studies have been successful or worthless’. Subsequently, Mercer (2010) provided a detailed review of ‘methods and methodologies’ for analysing classroom talk that considered the relative merits of a range of quantitative and qualitative methods. Experimental or quasi-experimental research offers the advantage of enabling the exploration of causal relationships between two variables. Such approaches have, therefore, frequently been seen as necessary for rigorous testing of the impact of educational interventions.

The lack of appropriate quantitative research has been seen as weakness in many of the earlier evaluations of Philosophy for Children. For example, Sternberg and Bhana voiced concerns over the lack of adequate control groups when they considered 20 evaluation studies of P4C many years ago. They noted that:

issues of subject dropout, class selection, durability, transfer, subject population, and experimenter bias were generally not addressed. Statistical analysis was often reported in only the most minimal detail: in some cases, significance levels were presented with no descriptive or inferential statistics.

(1996, p. 64)

Nevertheless, they noted widespread gains on verbal tests of critical thinking abilities. They also suggested that P4C was ‘highly teacher-sensitive and requires extensive teacher training’. Garcia-Moriyon, Robollo and Colom (2005) were also worried about the rigour of previous evaluations of Philosophy for Children, noting the wide variety of approaches and designs used (often brief subjective descriptions that were usually positive). Most reports lacked a clear description of the methodology used and sample size was sometimes very small.

The application of a predominantly positivistic approach typical of cognitive psychology to critically evaluate a philosophical process is not without contradictions in terms of the tension between ‘philosophical puzzlement and uncertainty’ and the ‘empirical truths’ of cognitive psychology. Some researchers (e.g. Robson, 1993) have advocated mixed methods according to what best fits a particular study. Mixed methods also have the advantage of increasing reliability by triangulating across methods. A mixed method approach involving both quantitative and qualitative methodologies was favoured by Topping and Trickey in their studies of the use of philosophical inquiry in Scottish classrooms.

Previous studies of Philosophy for Children

Meta-analyses

Meta-analytic processes are used to evaluate the effectiveness of a programme that has been subjected to a wide range of studies. The calculation of effect sizes enables the results from different
evaluation studies to be compared and combined into a quantitative index of effectiveness. Two relatively recent reviews of studies of the effectiveness of Philosophy for Children are those of the present authors (Trickey & Topping, 2004) and Garcia-Moriyon, Robollo and Colom (2005). Both analyses combined selected controlled evaluation studies of Philosophy for Children.

The present authors’ meta-analysis (Trickey & Topping, 2004) noted that although there was a great deal of descriptive literature, studies of outcomes were rather rare. Indeed, a similar state of affairs still pertains. This critical review included only studies using pre-post measurement of experimental and control or comparison groups. Ten studies met the stringent criteria for inclusion, measuring outcomes by norm-referenced tests of reading, reasoning, cognitive ability and other curriculum-related abilities, by measures of self-esteem and child behaviour, and by child and teacher questionnaires. All studies showed some positive outcomes. The mean effect size was 0.43 with low variance, indicating a consistent moderate positive effect for P4C on a wide range of outcome measures. Eight effect sizes were available from eight separate studies, ranging from 0.31 to 0.59, with a mean of 0.425 and a standard deviation of 0.09. Effect sizes of this magnitude are usually described as moderate and are certainly of educational significance. The low variance here is of interest.

The review showed a wide range of evidence for positive outcomes from different countries with different age groups of children. Many of the studies could be criticised on grounds of methodological rigour, but the quality and quantity of evidence nevertheless bears favourable comparison with that on many other methods in education. Not all measures within a study showed significant gains for experimental participants, but this perhaps reflects the difficulties of instrumentation sensitivity in complex real-life social environments. There were no negative findings in these controlled studies. However, it must be remembered that some of these studies were in contexts already committed to the use of P4C, and might not be typical.

Garcia-Moriyon, Robollo and Colom (2005) identified 18 studies of the impact of Philosophy for Children on reasoning skills. They computed a higher overall effect size of 0.58 but with ‘significant variability’. Their overall conclusion was that ‘children do improve their cognitive skills through this methodology’. Interestingly, Garcia-Moriyon et al. noted that their studies tended to take place for one school year, whereas Lipman designed P4C to take place across several years. In view of this, they speculated that gains would be greater with longer exposure to philosophical inquiry. The subsequent Scottish study set in Clackmannanshire schools (Topping & Trickey, 2007a) took place over a longer time frame and found an impact on reasoning ability with an effect size of 0.75. The Scottish study also provided evidence of generalisation across reasoning skills (verbal to nonverbal) and sustainability of impact beyond the initial intervention. This last finding is important and discussed in more detail later in this chapter.

**Recent evaluation studies**

Turning now to more recent studies, we find only two outcome studies, neither with follow-up. Lam (2012) studied 28 randomly selected first year secondary students in Hong Kong. They were randomly assigned to a group receiving P4C and another receiving English lessons. The P4C group had a significantly greater performance on a reasoning test and had a more positive attitude towards doing philosophy in the classroom. Ten crucial factors considered to contribute to the success of the P4C programme were outlined. This study was interesting for its cultural context, one where pedagogical methods tend to be prescriptive.

The second study (Reznitskaya et al., 2012) investigated whether the 11 year old participants improved their performance on various tasks measuring argumentation development. Twelve classrooms were randomly assigned to P4C and regular instruction conditions. The three post-intervention
measures were an interview, a persuasive essay and a recall of argumentative text. P4C students engaged in more dialogic interactions. However, they performed similarly to the regular instruction students on post-intervention measures.

Both of these studies lacked any kind of follow-up. One indicated gains in dialogue and in outcomes, while the other indicated gains in dialogue but not in outcomes. In the second study gains might have been expected in the interview, if not in the other measures.

Other studies focus more on the process of P4C. Daniel et al. (2005) investigated 10–12 year old pupils using P4C in relation to mathematics. Over a whole school year, transcripts of verbal exchanges revealed that dialogue was certainly established among the pupils, particularly in relation to logical, creative, responsible and meta-cognitive thinking. Reznitskaya and Glina (2013) examined the testimonials of 60 elementary school students about their experience during class discussions of assigned readings. Twelve classrooms were randomly assigned to P4C and regular instruction. Ten students from each classroom were interviewed. Significantly more P4C students said they enjoyed expressing disagreement with peers, taking on new responsibilities, and explaining their thinking to others. However, more P4C students complained about difficulties with getting the floor to speak, and suggested that better balanced group participation was needed.

Poor readers were targeted by Jenkins and Lyle (2010), albeit only four of them. The pupils were video recorded while engaged in discussion of questions they formulated themselves in response to a series of texts in preparation for a community of philosophical inquiry. Group discussions were analysed, paying attention to verbal and non-verbal communication. These poor readers did demonstrate their ability to use higher-order language skills. Gillies, Nichols and Burgh (2011) studied 18 sixth grade classrooms, involving 35 groups of students in three conditions: P4C, cognitive questioning and comparison, all in the context of science lessons. Students were videotaped as they worked on a specific inquiry-science task for two consecutive school terms. However, students in all conditions demonstrated more helping discourses or discourses known to mediate learning than any other of the discourse categories. This outcome is encouraging because it is the helping discourses where students provide explanations, elaborations and reasons that promote follow-up learning.

Thus these process studies generally claim to have identified enhanced dialogue and cognitive skills, but the comparative study of Gillies et al. (2011) warns that where a quasi-experimental investigation of process is undertaken, the comparison group can show gains as large as the intervention groups. One other study is of some relevance (Lizarraga, Ugarte, Iriarte and Baquedano, 2003), but it is a study of some elements of P4C mixed with aspects from other studies. It is encouraging in that it includes follow-up two years later, when positive outcomes were still evident, but it cannot be classed as a study of P4C. Forty 13 year old students were studied over two school years. Intervention included selected tasks from the Instrumental Enrichment Program, the Philosophy for Children Program, and Project Intelligence. Intervention was effective in the short term, and also at follow-up two years later.

Summary

The meta-analyses of P4C show interesting contrasts. Trickey and Topping (2004) found an effect size of 0.43 in 10 studies using a variety of measures with low variance, while Garcia-Moriyon et al. (2005) found an effect size of 0.58 purely in reasoning skills in 18 studies but with high variance. The Scottish studies found an effect size of 0.75 in reasoning among a mix of measures, with sustained gains at long-term follow-up. More recent literature is not all positive (although none is negative). Lam (2012) found short-term gains in reasoning. Daniel et al. (2005), Jenkins and Lyle
(2010), Reznitskaya et al. (2012) and Gillies et al. (2011) found improved dialogic reasoning during P4C classes, but Reznitskaya found no better outcomes and Gillies found process gains were equivalent in a control group. The only study to include long-term follow-up mixed elements of P4C with elements from other programmes (Lizarraga et al., 2003).

The present research

The study in Clackmannanshire, Scotland used Cleghorn’s (2002) development of Lipman’s programme – ‘Thinking through Philosophy’. The Thinking through Philosophy programme provided stories and poems with a high degree of ‘ambiguity’ to stimulate thinking and discussion. The research design involved both quantitative and qualitative methods. The evaluation consisted of three distinct methodologies providing a triangulated approach to the overall evaluation, that is:

- standardised tests administered to experimental and control classes to provide measures of cognitive ability (CAT3) and self-esteem (MALS);
- the analysis of classroom discussion using video recordings to provide observational measures of reasoning;
- analysis of the perceptions of children using questionnaires to provide an indicator of social outcomes.

The Cognitive Abilities Test (Lohman, Thorndike and Hagen, 1993) used multiple-choice questions and provided standardised scores of Verbal Ability, Nonverbal Ability and Quantitative Ability for each pupil. Pre-intervention standardised scores were obtained for 105 experimental pupils who went on to participate in regular inquiry lessons and 72 matched control pupils who followed their traditional curriculum. Follow-up testing took place 16 months later. There was an average gain of six standardised points in the Cognitive Abilities Test score of the experimental subjects. There were no gains in the cognitive ability scores of the control group. Interestingly, there were significant gains in all three cognitive ability areas (i.e. verbal, nonverbal and quantitative ability) in the experimental group, despite the intervention process being purely verbal. The difference between pre- and post-test CAT scores for the experimental pupils was found to be highly significant (p<0.001). The results thus suggest that one hour’s use of an inquiry-based teaching methodology each week can improve children’s reasoning ability (Topping & Trickey, 2007a).

Burden’s (2000) ‘Myself as a Learner’ Scale (MALS) was constructed to have a deliberate focus upon ‘academic self-concept’ as against all-embracing ‘self-concept’ or more specific academic-related scales (e.g. reading or mathematics self-concept). MALS provides a measure of students’ perceptions of themselves as learners and active problem solvers within educational settings. Pre-intervention scores were obtained for 186 pupils. This procedure was then repeated six months later as a post-test measure. Those pupils who had been involved in the Philosophy programme significantly improved their self-as-a-learner esteem scores (p<0.05). There was no difference between the pre- and post-test results of the control pupils. These results suggest that inquiry-based approaches can be conducive to promoting self-esteem in learning situations (Trickey & Topping, 2006).

Twelve video recordings of P4C in action (six pre-test and six post-test) were obtained, of classroom discussion of a researcher-specified Greek fable. The teacher first read out the story and then explored its meaning through discussion with the class. A number of ‘prompt questions’ were provided to help get the discussion started. Experimental classes then had a weekly
Philosophy lesson for six months, while control classes had no involvement. The teachers discussed the same Greek fable six months after the start of the inquiry lessons and the discussions were again video-recorded. The first 10 minutes of each video-recorded classroom discussion was scored using a structured observation schedule. Specific behaviours were selected for scoring on the grounds that they were readily observable and measurable and provided an indication of whether the broader aims of the programme had been achieved. A measure of inter-observer agreement was gained to ensure the observation schedule was sufficiently reliable. The experimental classes increased their scores in the targeted behaviours (all significant at $p<0.05$ level). There were no significant changes in the control classes. This provided evidence that philosophical classroom inquiry can result in increased participation of pupils, more elaborated responses from pupils, increased use of reasoning and justification by the children and an increase in the use of open-ended follow-up teacher questions (Topping & Trickey, 2007b).

The perceptions of pupils, teachers and head-teachers were obtained through analysis of questionnaires completed by 77 pupils and by head-teachers after six months. Verbal and written comments by participating teachers throughout the initiative were also compiled (elicited during support meetings and diaries maintained by the teachers). During analysis, all pupil responses for each open-ended question were assigned to categories that were judged to reflect clear themes in the pupils’ responses. The reliability of this approach was evaluated by establishing inter-rater reliability. Pupil responses to each question were subjected to chi-square analysis to determine the probability of the scores occurring by chance. The questionnaire responses indicated that the pupils perceived their involvement in ‘communities of inquiry’ as leading to an increase in their participation in classroom discussion and positive changes in their social behaviour and confidence in the classroom (Trickey & Topping, 2007).

Evidence of positive changes in pupil social behaviour were strengthened by the consistency of pupil responses to the different questions in the questionnaire; the consistency between pupil perceptions, class-teacher perceptions and head-teacher perceptions; and the consistency of findings from different evaluation methodologies. One example of consistency of findings from different evaluation methodologies was that the questionnaires and the video analysis of classroom discussion both provided clear evidence of increased participation of pupils in classroom discussion. Another was that pupil perceptions of increased confidence were consistent with findings from the MALS measure of academic self-esteem. In other words, outcomes were consistent across different evaluation methodologies.

**Long-term effects of Philosophy for Children**

The time and effort that teachers invest in the use of philosophical inquiry in the classroom are only going to be merited if there are long-term gains from such interventions. However, very few studies of the effectiveness of Philosophy for Children have included short- or long-term follow-up. Short-term positive effects may wash out. Conversely, it is also possible that there are ‘sleeper’ gains that could take years to show. Investigators need to be mindful of Hawthorne effects arising with a novel and enthusiastically introduced programme. These may heighten initial positive outcomes, but gains may prove difficult to generalise to other contexts or sustain in the longer term. There has been evidence of long-term gains from other programmes: Cognitive Acceleration through Science Education (Adey and Shayer, 1994) and Feuerstein’s Instrumental Enrichment Program (Feuerstein, Rand, Hoffman and Miller, 1980). However, overall such evidence remains sparse, particularly in relation to Philosophy for Children.

In view of this, the Scottish study investigated the cognitive effects of collaborative philosophical inquiry at long-term two-year follow-up, after the participants had transferred to secondary
(high) school without experiencing further philosophical inquiry in the interim (Topping and Trickey, 2007c). Intervened children had engaged in collaborative inquiry for one hour per week over 16 months in the primary (elementary) school. Sample attrition was greater in the control than in the experimental group, but 96 experimental and 52 control subjects were available. The Cognitive Abilities Test was again used. The significant pre-post cognitive ability gains in the experimental group in primary school were maintained. Higher achieving pupils were somewhat advantaged in sustaining these gains. The control group showed an insignificant but persistent deterioration in scores from pre- to post-test to follow-up. The study provides evidence of maintained cognitive gains beyond the initial intervention.

**Implications for practice, policy and future research**

The finding of long-term gains (Topping and Trickey, 2007c) was significant in view of the rarity of follow-up studies of the effects of collaborative philosophical inquiry. It suggests that while cognitive gains resulting from regular classroom inquiry may be gradual (as seen by both teachers and school administrators), when they do occur they may be of long-term duration. There was also unusual evidence of reasoning abilities being generalised from verbal reasoning to non-verbal and quantitative reasoning abilities. Both these observations of long-term gains and generalisability are consistent with the findings of Adey and Shayer (1994), where initial gains were reflected in academic results three years later and where gains in one subject (Science) generalised to other subjects (English and Mathematics). Cognitive interventions based on promoting structured classroom dialogue may thus merit the investment of valuable teaching time in the classroom and both pre- and in-service professional development opportunities for teachers. However, the quality of training and on-going support for teachers will be two of many potential confounding factors in final outcomes.

From a policy perspective, the indication of long-term gains assumes additional significance in view of the correlation between cognitive reasoning abilities and academic achievement (e.g. Deary, Strand, Smith and Fernandes, 2007), the need for critical thinking abilities in the workplace (e.g. Organisation for Economic and Co-operative Development cited in Powney & Lowden, 2000) and the need for critical thinking in a healthy democracy (Abrami et al., 2008). All of these are reasons why government policy would do well to move away from narrow compartmentalised subject specialisms and towards a more integrative ability to reason and apply this reasoning to real life – in short, to think.

For future research, the implications are relatively clear. Mixed methods are likely to offer more substantive findings than either a narrow quantitative or qualitative approach. Quasi-experimental designs with pre- and post-tests would seem to be an essential component of such approaches. Long-term follow-up is essential if sustained gains are to be found. If this follow-up occurs after transfer to a new kind of school, so much the better.

**Conclusion**

The average effect sizes of meta-analyses vary between 0.43 and 0.58, and no studies report negative effects. More recently, there has been another study of short-term reasoning gains, and several of the dialogic processes deemed important in P4C. However, not all results clearly favoured P4C. One study found long-term gains but did not only feature P4C. The Scottish studies remain the most unequivocally positive, with substantial gains in comparison to controls (effect size 0.75) which endured for two years in a different school after P4C had ended. From a practitioner perspective, it seems that P4C is worth implementing, given its effects, the amount
of curricular time it displaces and that it is recordable under both language, social, emotional and possibly other curricular headings. From a policy perspective, long-term enhancement of transferable thinking skills has to be good not only for schooling but also for subsequent economic life, not to mention quality of well-being. From a research perspective, more mixed-methods research involving follow-up is necessary. However, we also need to explore the process of P4C, to establish what aspects are crucial to gains and to render implementation sturdy in the face of larger scale-up.

References


Mercer, N. (2010). The analysis of classroom talk: Methods and methodologies. British Journal of Educational Psychology, 80, 1–14.


