British Journal of Educational Psychology (2007), 77, 271–288

The British Psychological Society

www.bpsjournals.co.uk

Collaborative philosophical enquiry for school children: Cognitive effects at 10–12 years

K. J. Topping[1]* and S. Trickey[2]

[1]University of Dundee, Scotland, UK

[2]Clackmannanshire Education Authority, Scotland, UK

Background. Debates about the modifiability of cognitive ability have been largely resolved by reports of successful 'thinking skills' interventions. However, such interventions are very diverse and generalization of effects relatively little explored.

Aims. This study investigated whether a thinking skills intervention involving collaborative interactive dialogue could lead not only to gains in measured verbal cognitive ability but also generalization to non-verbal and quantitative reasoning ability.

Sample. Randomly selected intervention children were aged 10 at pre-test ($N = 105$, four classes/schools). Controls followed a normal curriculum ($N = 72$, three classes/two schools).

Method. Intervention children engaged in collaborative enquiry for 1 hour per week over 16 months. The control group received normal classroom experiences. The Cognitive Abilities Test was administered before and after the intervention.

Results. Intervention pupils showed significant standardized gains in verbal and also in non-verbal and quantitative aspects of reasoning, consistent across intervention schools. Boys and girls made significant gains. The highest quartile of pre-test ability showed the smallest gains. Controls did not gain in any aspect.

Conclusions. Philosophical enquiry involving interactive dialogue led not only to significant gains in measured verbal cognitive ability but also generalization to non-verbal and quantitative reasoning ability, consistent across schools and largely irrespective of pupil gender and ability. The effect sizes from this large-scale field trial in one local authority exceeded those reported in the literature. Implications for theory building, replicability and sustainability are addressed.

The notion of intelligence as a fixed non-modifiable ability has largely fallen into disrepute, although as recently as 1994 Hernstein and Murray contended that intelligence was very difficult to change. Sternberg and Jensen (1992) and Dweck (2000) reviewed and contrasted theories of fixed intelligence (entity theories) and

* *Correspondence should be addressed to Keith Topping, Professor of Educational & Social Research, School of Education, University of Dundee, Gardyne Road, Dundee DD5 1NY, Scotland, UK (e-mail: k.j.topping@dundee.ac.uk).*

DOI:10.1348/000709906X105328

malleable intelligence (incremental theories) – and how beliefs about these could affect pupil performance.

In parallel, the conception of intelligence as a single variable (the 'g' factor) has been repeatedly challenged in recent decades, by Sternberg for example (Sternberg, 1985, 1997), and more recently by the popularization of notions such as 'multiple intelligences' (Gardner, 1999) and 'emotional intelligence' (Goleman, 2005). However, the empirical support for these latter is limited, and such models of intelligence have been extensively critiqued in reviews (e.g. Matthews, Zeidner, & Roberts, 2002). If multiple intelligence factors are manifested in situated contexts (Cobb & Bowers, 1999), little transfer between factors might be expected to result from an intervention involving mainly one factor.

The present study was a test of the incremental theory, but more particularly a test of whether an intervention delivered primarily through verbal dialogue could have generalized effects upon non-verbal aspects of intelligence. It also explored the influence on cognitive outcomes of school diversity and pupil characteristics, such as gender and ability. It was also a test of scalability of cognitive modifiability; i.e. whether a large-scale field implementation of an intervention in intelligence could consistently yield results as positive as those reported from research studies in the literature – at affordable resource costs.

Cognitive interventions

A number of examples of successful interventions in intelligence have been reported (see reviews by Adey & Shayer, 1994; Cotton, 2002; McGuinness, 1999; Moseley *et al.*, 2004; Sternberg & Bhana, 1996; Thinking Skills Review Group, 2004; Wilson, 2000). However, many of these cognitive interventions were intensive, long-lasting and costly. Some involved high teacher/pupil ratios (e.g. Feuerstein, Rand, Hoffman, & Miller, 1980), with implications for cost effectiveness, sustainability and replicability. Investigation of intervention effectiveness with normal class sizes is therefore of interest.

Moseley *et al.* (2004) classified frameworks for thinking skills under four headings: all-embracing frameworks covering personality, thought and learning; instructional design frameworks; frameworks for developing critical and productive thinking; and explanatory models of cognitive structure and/or cognitive development. Given this variation, programmes were likely to further vary in:

- Cognitive focus, i.e. the extent to which verbal, non-verbal or other kinds of thinking are engaged and targeted.
- Interactivity, i.e. the extent to which they rely upon individual work and/or teacher direction (often based upon specified materials), or by contrast upon some form of cooperative or collaborative learning (flexibly supported by the teacher).
- Generalization potential, i.e. the probability of generalization across subtypes of intelligence and the curriculum and beyond
- Opportunity costs (i.e. the extent to which they displace other curricular activity) and cost effectiveness, which is likely to affect sustainability.

These last two issues have rarely been addressed in the literature. However, Wegerif, Mercer, and Dawes (1999) showed gains on Raven's nonverbal reasoning test apparently stemming from their Exploratory Talk programme. Adey and Shayer's (1994) CASE programme is delivered in science but reported to yield outcomes in mathematics and English language public examination results. Shayer and Adey

(2002) report extension of this work in which the delivery curriculum subject is itself mobile.

The intervention in the current study (Philosophy for Children) was a 'critical and productive' thinking programme, characterized by a verbal cognitive focus and high self-regulated peer interactivity developing from initial teacher scaffolding. It operated only for 1 hour per week, and modest generalization potential and opportunity costs might be presumed. Within Gardner's (1999) conception of multiple intelligences, gains might be expected in linguistic intelligence, but presumably not logical-mathematical intelligence or spatial intelligence (or most of the multiple intelligences).

Language skills are related to socio-economic factors (e.g. Locke, Ginsberg, & Peers, 2002), and so generalization of effects from verbally loaded interventions is important if such interventions are not to favour more socio-economically advantaged children. Indeed, one outcome might be that the gap between the more and less intelligent widens. Given the diversity of programmes, are some more suited to female learners, risking increasing the attainment gap between girls and boys? Some interventions are particularly complex and require sophisticated teacher competencies and high organizational reliability, thus durability of effectiveness across diverse schools is also of interest.

The Thinking Skills Review Group (2004) commented: 'Further work is needed on identifying efficient, as well as effective, ways of intervening to promote thinking skills and raise attainment' (p. 5). Additionally, school and pupil characteristics and aptitude × treatment variation have not been systematically investigated. The present study sought to explore some of these concerns, with respect to the intervention 'Philosophy for Children'.

Theoretical modelling

But why *should* there be generalization from verbal interaction to other areas of cognitive functioning? Reviews of the literature on transfer or generalization of learning (e.g. Campione, Shapiro, & Brown, 1995; Perkins & Salomon, 1994; Singley & Anderson, 1989) emphasize: questioning; meta-cognitive strategy development, self-regulation and self-esteem; constructing knowledge within a social environment (cf. Wenger, Pea, Brown, & Heath, 1999); the development of learner self-monitoring of strategies and their effectiveness; and in addition, emphasis on the structural similarities of diverse problems and the direct teaching and multiple exemplification of generalizable principles and concepts.

More specifically, Topping and Ehly (2001) proposed a theoretical model of peer-assisted learning in which cognitively demanding peer interactions could include the following elements: individualizing goal and plans; peer modelling, self-disclosure and accountability; hypothesizing, questioning, explaining, clarifying, simplifying, rehearsing, prompting, summarizing; error detection, diagnosis and resolution. In a process of co-construction, these had potential to enhance meta-cognition, self-monitoring and self-regulation of learning, with consequential self-attribution of learning success and thereby self-esteem as a learner. Of these elements, peer modelling, peer self-disclosure, hypothesizing, questioning, meta-cognitive strategy development, self-regulation, self-attribution for success and self-esteem as a learner appear to have potential for enhancing generalization.

Cognitive interventions featuring a larger number of these desirable elements from both literatures might be more likely to demonstrate generalization.

Philosophy for children

Philosophy for Children (Lipman, 1981, 1991, 2003; Lipman & Sharp, 1978; Lipman, Sharp, & Oscanyon, 1980) has often been thought of as a separate programme (e.g. McGuinness, 1999). However, the method of Philosophy for Children (P4C) can be infused into a range of subject domains (Fisher, 1999), although it is usually initiated as a separate activity. Lipman strongly espoused the principle of cognitive modifiability (2003, p. 40) and advocated 'converting the classroom into a community of enquiry' (1991, p. 15).

Lipman's writing tends to be dynamic and illustrative rather than fixed and structural (e.g. 1991, 2003), but he conceptualized three modes of thinking (*critical*, *creative* and *caring*) and four main varieties of cognitive skill (*enquiry*, *reasoning*, *concept formation* and *translation*). Critical thinking is sensitive to context, relies on criteria and has high potential for self-correction. Creative thinking is imaginative, holistic, inventive and generative. Caring thinking is appreciative, normative, affective and empathetic. Enquiry is a self-corrective practice in which a subject matter is investigated with the aim of discovering or inventing ways of dealing with what is problematic, the products of which are judgments. Reasoning is the process of ordering and coordinating what has been found out through the enquiry. It involves finding valid ways of extending and organizing what has been discovered or invented while retaining its truth. Concept formation involves organizing information into relational clusters and then analysing and clarifying them so as to expedite their employment in understanding and judging. Conceptual thinking involves the relating of concepts to one another so as to form principles, criteria, arguments and explanations. All of these elements constitute Lipman's definition of collaborative philosophical enquiry, and featured in the intervention reported below.

P4C incorporates features that Adey (2001) has suggested are critical for promoting cognitive skills and educational attainment, including the verbal dialogue seen as important in Piagetian and neo-Piagetian theories of cognitive conflict and consequent assimilation and accommodation (e.g. Doise & Mugny, 1984; Piaget, 1932), which Mercer (2000) and Carnell and Lodge (2002) considered essential for rich learning environments. P4C also encourages children to become more aware of thinking and learning in themselves and others – likely to enhance meta-cognitive processes (Watkin, 2001).

In a systematic review of controlled outcome studies of 'Philosophy for Children' in primary (elementary) and high schools (Trickey & Topping, 2004), 10 studies met the stringent criteria for inclusion. These measured 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 (Cohen's δ) was 0.43 with low variance, indicating a consistent moderate positive effect for P4C on a wide range of outcome measures.

The 'Thinking through Philosophy' intervention

The intervention in this study was based on Lipman *et al.*'s (1980) 'Philosophy for Children' process, but used more contemporary practical programme materials – 'Thinking through Philosophy' (Cleghorn, 2002). Haynes (2001) summarized the operational process of philosophical enquiry in nine steps:

(1) Getting started – agreeing rules of interaction and beginning with a relaxation exercise.
(2) Sharing a stimulus to prompt enquiry.
(3) Pause for individual thought.
(4) Questioning – the pupils think of interesting or puzzling questions.
(5) Connections – making links between the questions.
(6) Choosing a question to begin an enquiry.
(7) Building on each other's ideas, during which the teacher has to strike a balance between encouraging the children to follow on from each other's ideas and allowing related lines of enquiry to open up.
(8) Recording discussion – graphic mapping.
(9) Closure and review – summarizing, reflecting on the process itself, considering whether minds were changed.

Cleghorn's (2002) 'Thinking through Philosophy' programme incorporates these elements, but describes stages in each lesson thus:

(1) Focusing exercise – scripted and aimed to create an alert but relaxed state in which the children's attention is more 'in the present'.
(2) Linking with the previous week – to reinforce memory of what has taken place the previous session and provide an opportunity to bring forward any new related thinking.
(3) Stimulus – typically a story or poem is read aloud by the teacher, accompanied by any relevant visuals.
(4) Pair work – providing an opportunity to check the children's initial understanding of the stimulus.
(5) Dialogue in groups of about six children – the teacher encouraging pupils to: (a) communicate their views in response to an agreed subject of enquiry; (b) support their views with reasons; (c) listen respectfully to views being expressed; (d) indicate whether they agree or disagree with those views; (e) provide alternative viewpoints; (f) gradually develop a process of dialogue that helps the group construct a deeper understanding (or better solution) than would be possible individually.
(6) Closure – encouraging children to reflect on the discussion and how their thinking might have progressed.
(7) Thought for the week – highlighting a practical idea drawn from the stimulus to provide homework for the rest of the week to help relate the main ideas to real situations outside that stimulus.

A key element is the emphasis on developing a community approach to enquiry in the classroom, characterized by open-ended Socratic questioning by the teacher, challenging the children to think more independently. Such questions are also instrumental in promoting teacher–pupil and pupil–pupil reciprocal dialogue.

For this intervention, in-service professional development support was coordinated by Cleghorn, head teacher of a local primary school, together with two senior teachers experienced in leading classroom enquiry. During the period of the evaluation, the combined time allocation of these three teachers to the initiative amounted to the equivalent of 0.2 of a full-time teacher. Intervention teachers from the last two primary years (pupils leave primary school at the age of 12 in this area) received a total of 10–12

hours of professional development during the first year of the initiative. Pre-intervention, this included a full day input plus observation of expert classroom practitioners and debriefing with that practitioner. Subsequently, each trimester participant teachers attended a 2-hour after-school group professional development session to share progress and talk through issues arising. The teachers thus formed their own community of enquiry. Further support was available from the specialist teachers on a call-out basis.

Elements of P4C fostering generalization

Considering both Lipman's general conceptualization of P4C and its operation in this study, many of the elements suggested by theoretical modelling as likely to foster generalization seem to be present. Investigation involves hypothesizing and questioning. Enquiry involves developing problem-solving skills and strategies. Knowledge is constructed in a social environment involving peer modelling and disclosure, with interaction rules to protect and enhance participant self-esteem. Reasoning involves ordering and coordinating what has been discovered, discerning the structural similarities of diverse problems and developing translational skill in extending this relational conceptualization to new contexts. The summarization and process feedback component is designed to promote meta-cognition and enhance self-regulation. The only desirable element missing is the direct teaching of generalizable concepts. It seems that the generative, extending, relational and strategic characteristics of P4C can lead to plausible predictions of generalization to areas of cognitive functioning beyond the verbal. However does this actually happen?

Aims

This study investigated whether a weekly collaborative enquiry intervention over time would lead to:

(1) Larger gains in measured cognitive ability than non-intervention controls;
(2) Gains in measured verbal cognitive ability, but also generalization to non-verbal and quantitative reasoning ability;
(3) Gains in measured cognitive ability irrespective of school attended, and pupil characteristics such as ability and gender;
(4) Gains in measured cognitive ability from a large-scale field trial equivalent to those found in research studies.

Method

Research design

The research design was a 2 × 2 pre-post intervention/control design, which could be considered quasi-experimental as sampling was not totally random and not all school/class effects could be controlled. Both groups were tested and retested under the same conditions. The pre-post period of 16 months took into account that gains from cognitive development interventions are usually expected to be gradual (e.g. the CASE study, Adey & Shayer, 1994). The longer period might also reduce practice and Hawthorne effects.

The authors felt that a design exploring causality and using a norm-referenced measure of generalized cognitive abilities would yield robust conclusions that were

widely acceptable. This is not to say that process factors were overlooked, and cognitive measures were triangulated with observations of interactions (reported elsewhere – Topping & Trickey, in press – and see Discussion below). Burden and Nicholls (2000) have argued that process and qualitative measures should take precedence in studies of thinking skills, but in the current project mixed methods were considered to be more appropriate.

Context and sample

The study was located in a small ethnically homogeneous local education authority (school district) of mixed socio-economic status but including pockets of severe disadvantage. The authority refused to accept totally random sampling for this study. All class teachers from the last 2 years of primary school were invited to a meeting where attendees were invited to participate in the first phase of the intervention. All 19 schools expressed interest, but existing commitments to school development plans left 8 schools that were free to engage with the first phase. The eight initial intervention schools were thus to some degree self-selected, but not on the basis of highest motivation or best organization.

The evaluation included other assessments in addition to cognitive ability. To make workload and curriculum intrusion acceptable, all intervention schools did not experience all forms of measurement. For cognitive abilities testing, four of the eight intervention schools and four classes from their eight intervention classes were randomly selected by drawing lots at both levels (the pupils aged 10 at cognitive abilities pre-test).

From the 11 non-intervention schools, four comparison or control schools (who had of course also expressed interest in participating in future intervention phases) were selected as comparable in terms of best fit in relation to the factors of pupil ability, size of school and social disadvantage. From these, a pool of matched classes was formed in terms of best fit to intervention classes in relation to the factors of pupil age and ability (but not gender) at the class level. From this pool, three classes were randomly drawn for cognitive abilities testing (in the event, from two control schools). The control children received classroom experiences and the class teachers continuing professional development experiences, normal for the local authority (school district) during the same time as the intervention. All control schools and classes subsequently became involved in the second phase of implementation.

This matched and random selection yielded an intervention group with a preponderance of males (male $n = 60$–57%, female $= 45$) compared with the control group (male $n = 27$–38%, female $= 45$). The pre-test mean standardized cognitive ability score was 99 for the intervention pupils ($SD = 12.1$) and 101 for the control pupils ($SD = 11.1$), both normally distributed. The mean roll (total number of pupils) for the intervention schools was 328, for control schools 365. The average number of children receiving free school meals (as an indicator of socio-economic disadvantage) was 52 in intervention schools and 61 in control. Only pupils for whom complete pre-test and post-test data were available were included in the study. Fortunately, it was undertaken in an area of low geographical mobility, and strenuous attempts were made to collect data on pupils absent at regular test sessions, so attrition was not significant.

Measures

The multiple-choice Cognitive Abilities Test (CAT) (Lohman, Thorndike, & Hagen, 1993) was selected as the preferred measure, not least as scores relate strongly to subsequent pupil attainment in external examinations at 16 years of age (Smith, Fernandes, & Strand, 2001), an outcome of great importance to children, families, teachers and the local authority. The updated version (CAT3) (Smith *et al.*, 2001) was used, just before commencement of the intervention and 16 months later (after which the intervention continued). Instructions are clearly scripted and the test is designed to be administered by class teachers. Class teachers for each year administered the test to whole classes. Pupil response sheets were dispatched to the test suppliers for machine scoring.

CAT3 was standardized on 15,859 pupils. In addition to an overall total standardized score, it provides scores for a total of nine subtests for each pupil, with reliability coefficients between 0.89 and 0.96. The test's validity was established through a factor analysis of the nine subtests and by correlation between CAT3 scores and other evidence of intellectual ability.

The nine CAT subtests encompass:

(1) Verbal classification test requires the selection of a word from five choices that conceptually relates to three other words (such as green, blue, red).
(2) Sentence completion test requires selecting a word from five to complete a sentence that has a word missing.
(3) Verbal analogies test provides two words that relate (e.g. new and old) – a third word that similarly relates must be selected from five choices.
(4) Number analogies test provides two sets of numbers that are related in some way (e.g. 2/3 and 6/7), and another number on its own. A number must be selected from a choice of five that relates to the number on its own.
(5) Number series test provides a series of numbers that are linked by a rule that needs to be deduced. The correct number must then be selected from a choice of five to continue the series.
(6) Equation building test provides three numbers and signs that can be combined to make answers. The correct combination must be selected from a choice of five.
(7) Figure classification test provides three figures that are related in some way. The relationship has to be deduced and a related selection made from five options.
(8) Figure analogies test provides two figures that are related in some way (e.g. two squares). A third figure is provided and the correct choice from five options must be made showing the analogous relationship.
(9) Figure analysis test shows a square of folded dark paper with holes punched into it. A choice showing how the paper would look when unfolded must be selected from five options.

These nine subtests are grouped into three aggregated subscales yielding standardized scores of verbal ability, non-verbal ability and quantitative ability (1–3, 4–6, 7–9 above). (The authors will send further details of CAT to enquirers on request.)

Analysis

It was considered that sampling constraints did not seriously contra-indicate the use of parametric statistical analysis. For single comparisons, related and unrelated *t* tests were used. For multiple comparisons, one-, two- and three-way mixed analyses of variance

were conducted to explore interactions between pre-post gains, group conditions, gender and school, followed where relevant by appropriately conservative *post hoc* tests in relation to homogeneity of variance. For single comparisons, effect sizes (ES) were calculated using Cohen's δ (experimental mean gain – control mean gain, standardized by control gain standard deviation), and for multiple comparisons ES was derived from ANOVA as partial eta-squared (η^2).

Results

Intervention and control gains

Pre-post data for the 105 intervention pupils are summarized in Table 1, with *t* test probabilities. There was an overall mean positive change of 6.0 standardized points ($SD = 6.7$). One-way within-subject repeated measures ANOVA on the total CAT standardized scores of the intervention group indicated a highly significant gain ($F(1, 104) = 69.274$, $p < .001$, $\eta^2 = .449$). The largest average gain was in non-verbal ability (7.2 points); the most modest in quantitative ability (5.0 points).

Table 1. Pre-post standardized scores for intervention ($N = 105$) and control ($N = 72$) groups

	Pre-test		Post-test		Change		
CAT subscale	Mean	*SD*	Mean	*SD*	Mean	*SD*	Probability
			Intervention				
Verbal	99.0	13.2	104.8	13.3	5.8	13.3	<.01
Quantitative	99.0	14.8	104.0	15.4	5.0	15.4	<.01
Non-verbal	99.0	14.6	106.2	13.6	7.2	13.6	<.01
Overall	99.0	13.1	105.0	14.1	6.0	6.7	<.01
			Control				
Verbal	99.7	12.6	99.0	15.2	−0.7	15.2	.69
Quantitative	101.6	11.2	99.0	12.0	−2.6	12.0	.11
Non-verbal	102.8	12.2	100.2	12.4	−2.6	12.4	.20
Overall	101.3	12.0	99.4	13.2	−0.9	13.2	.33

Pre-post data for the 72 control pupils are also summarized in Table 1. There were no significant changes in the scores of the control group in any area. Post-test scores tended to be lower than pre-test.

A two-way mixed ANOVA with 'group' (intervention or control) as a between-subjects factor and 'pre-post' (pre-test or post-test) as a within-subjects factor showed a large significant effect for the interaction 'pre-post × group': $F(1, 175) = 49.516$, $p < .001$, $\eta^2 = .267$. The main effect of the factor 'group' was very small and far from significant: $F(1, 175) = 0.086$, $p = .770$. This suggests that there were no differences between groups other than that resulting from the intervention.

Effects on verbal, non-verbal and quantitative abilities

Differences in gains on the three CAT subscales were then analysed separately. On the verbal subscale, a two-way mixed ANOVA pre-post × group showed a significant advantage for the intervention group: $F(1, 175) = 20.911$, $p < .001$, $\eta^2 = .131$. On the non-verbal subscale, a two-way mixed ANOVA pre-post × group showed a significant

advantage for the intervention group: $F(1, 175) = 23.276, p < .001, \eta^2 = .146$. On the quantitative subscale, a two-way mixed ANOVA pre-post × group showed a significant advantage for the intervention group: $F(1, 175) = 12.414, p = .001, \eta^2 = .084$. Thus all three subscales showed significant intervention gains.

Consistency across schools

Overall CAT scores for the intervention group were then analyzed by school (Table 2). All the intervention classes increased their mean overall cognitive ability scores significantly from pre-test to post-test. A mixed 2 × 7 ANOVA pre-post (pre-test or post-test) × school across all schools showed no significant pre-post × school interaction: $F(5, 170) = 1.632, p < .156$. *Post hoc* tests (Bonferroni) indicated that one intervention school showed significantly worse results than other intervention schools (there was no evidence that this was plausibly attributable to lower socio-economic status), but otherwise few differences emerged. This indicated a high level of consistency in achieving favourable outcomes.

Table 2. Overall CAT standard score mean gains in intervention schools/classes

	School A	School B	School C	School D
Number of pupils (*N*)	15	30	29	31
Pre-test score	103.6	101.6	97.0	96.1
SD	17.3	10.8	10.9	14.2
Post-test score	111.3	107.6	101.2	103.5
SD	15.0	8.8	11.8	13.0
Gain	7.7	7.0	4.2	7.4
SD	5.1	5.8	5.6	6.5
Probability	<.01	<.01	<.01	<.01

Table 3 provides a breakdown by intervention school of the subscale results (verbal, quantitative and non verbal measures of cognitive ability).

All schools showed gains on all measures. Many schools did well on both verbal and non-verbal subscales. However, the gains of two schools in the quantitative reasoning area did not reach significance. Some variation between schools was evident, one school doing especially well on the verbal subscale and another on the quantitative subscale.

Size of the class did not appear a factor in gains. Three of the classes averaged 30 pupils and one had half this number of pupils. The highest increase was in the smallest group but only just. Two of the three large classes also achieved gains of seven standard points or more.

Table 4 shows the comparable probabilities (*t* test) for the intervention and control schools. None of the latter was statistically significant (results from the two control classes in one school were similar and are aggregated).

Effect of pupil pre-test ability

Intervention schools with mean overall pre-test scores above average and below average gained significantly. However, when the whole intervention group was split into quartiles according to pre-test score (pre-test score <91, 91–98, 99–106, >106) (Table 5), the highest ability quartile had by far the smallest gain. The two middle

Table 3. CAT subscale standard score mean gains in intervention schools

	School A	School B	School C	School D
Number of pupils (*N*)	15	30	29	31
Verbal pre-test	106.2	102.1	96.2	95.2
SD	18.1	11.5	11.1	12.7
Verbal post-test	110.9	106.6	101.2	102.7
SD	15.9	12.4	11.0	13.7
Verbal gain	4.7	4.5	5.0	7.5
SD	4.3	6.1	11.0	4.6
Probability	.03	.03	.01	<.01
Quantitative pre-test	103.8	100.9	97.2	96.6
SD	14.9	12.1	10.3	17.1
Quantitative post-test	110.6	105.7	101.2	101.2
SD	16.6	11.2	12.5	19.9
Quantitative gain	6.8	4.8	4.0	4.6
SD	7.4	6.4	4.8	10.2
Probability	.05	.13	.03	.25
Non-verbal pre-test	100.7	102.0	97.5	96.7
SD	17.3	13.6	14.6	15.5
Non-verbal post-test	112.1	109.7	101.6	104.0
SD	15.7	10.4	12.3	13.3
Non-verbal gain	11.4	7.7	4.1	7.3
SD	8.3	11.2	6.0	5.8
Probability	.01	.01	.01	<.01

Table 4. Probabilities of subscale score change in intervention and control schools

	Verbal	Non-verbal	Quantitative
Intervention schools			
School A	.03	.01	.05
School B	.03	.01	.13
School C	.01	.01	.03
School D	<.01	<.01	.25
Total	<.01	<.01	<.01
Control schools			
School E	.32	.51	.75
School F	.23	.08	.09
Total	.69	.11	.21

quartiles showed the highest gains, with the lowest quartile not far behind (despite being depressed by the one intervention school with the lowest gains). A mixed 2×4 ANOVA was conducted (within-subjects pre-test and post-test score, between-subjects quartile group). This indicated a significant pre-post $\times$ quartile interaction, suggesting that some quartile groups benefited from the intervention more than others: $F(3, 101) = 4.010$, $p = .010$, $\eta^2 = .135$. As quartile error variances were not equal at pre-test, Games-Howell *post hoc* tests were conducted, which indicated that every quartile was significantly different from every other quartile (all $p < .001$). There was little evidence of regression to the mean.

Table 5. Intervention pre-post scores by quartile of pre-test score

		Post-test score		Post-test score		Gain	
Quartiles	*N*	Mean	*SD*	Mean	*SD*	Mean	*SD*
First quartile	26	83.53	4.695	89.16	8.821	5.63	6.784
Second quartile	26	94.95	1.545	102.26	6.217	7.32	6.210
Third quartile	26	102.05	2.915	108.16	4.799	6.11	4.852
Fourth quartile	27	116.12	7.514	117.92	7.940	1.79	4.587
Total	105	100.21	13.157	105.21	12.778	5.21	5.608

Effect of pupil gender

The intervention group cognitive ability overall and subscale scores were analysed by gender (Table 6). Both male and female gains achieved statistical significance ($p < .02$). A three-way mixed ANOVA on overall pre-post (pre-test or post-test) × gender (male or female) × group (intervention or control) showed no significant differences in gains by gender: Pre-post × gender: $F(1, 173) = 0.009$, $p = .924$; pre-post × group × gender: $F(1, 173) = 0.558$, $p = .457$. Boys' pre-test scores were slightly higher than those of girls on all subscales. Boys made higher gains than girls in overall scores and in verbal and non-verbal subscale scores. Girls made higher gains than boys in quantitative subscale scores.

Table 6. Intervention CAT overall and subscale mean scores by gender

	Male	Female
N	60	45
Pre-test overall mean	99.4	98.4
Post-test overall mean	106.5	103.5
Gain	7.1	5.1
SD of gain	6.7	5.9
Probability	<.01	<.01
Pre-test verbal mean	99.9	97.8
Post-test verbal mean	106.4	102.4
Gain	6.5	4.6
SD of gain	4.6	5.9
Probability	<.01	<.01
Pre-test quantitative mean	99.2	98.8
Post-test quantitative mean	103.6	104.7
Gain	4.4	5.9
SD of gain	7.4	7.7
Probability	.01	.01
Pre-test non-verbal mean	99.3	98.6
Post-test non-verbal mean	108.2	103.5
Gain	8.9	4.9
SD of gain	6.9	9.2
Probability	<.01	.02

Effect sizes

For overall CAT scores the ES (δ) was 0.75. For non-verbal CAT scores the ES (δ) was 0.79. For verbal CAT scores the ES (δ) was 0.73. For quantitative CAT scores the ES (δ)

was 0.69. These effect sizes compare very favourably with the average ES (δ) of 0.43 with low variance derived from 10 previous research studies of P4C by Trickey and Topping (2004). ESs derived from analyses of variance (η^2) were more modest.

Discussion

This study reports good news – that P4C yielded cognitive gains compared with controls that transferred across domains of intelligence, were largely irrespective of pupil school/class, pre-intervention ability and gender, and which cost relatively little to achieve. However, these welcome results should not be over-interpreted or accepted uncritically.

Methodological issues

Obviously this study had imperfections. Sampling was not entirely random (although there was little evidence of bias in the sample, and totally random sampling might have affected external validity). Intervention and control groups did have different gender proportionalities (intervention more male, control more female), but it is debatable whether this strengthens or weakens the claim for greater effects for males. The study relied on a single measure of cognitive ability (albeit one of established reliability and validity, known to correlate highly with subsequent educational achievement). The administration of the testing was standardized as far as possible. The test used was designed to be administered by class teachers and provided clearly scripted instructions for the teachers. The test publisher scored the answer sheets by computer, ensuring objectivity in scoring. The possibility of Hawthorne effect in the intervention gains must be countenanced, since P4C was considerably different to previous teaching and might have had a novelty and/or inspirational effect, especially as it was coupled with special professional development activities and support services. However, these latter were relatively lightweight, any sense of elite was unlikely as half the primary schools in the authority were involved and all would be over time, and it seems improbable that any Hawthorne effect would endure over 16 months of regular sessions of the same type. Despite the attempt to control for schools, it is possible that differences between experimental and control classes were influenced by background systematic teacher, class or school effects that were not measured. Effect sizes calculated from gain scores may overestimate effects compared with those calculated from post-scores where pre-test scores are identical for intervention and control groups.

Process factors

Measures of implementation integrity of the intervention were made by video recording and analysis of classroom interactions. These are reported elsewhere (Topping & Trickey, in press), but are summarized here to give the present reader a fuller picture. The quantity and quality of interactive dialogue in 180 children aged 10 in four P4C intervention and two control classes in six schools were studied. Video recordings of classroom discussions before, and 7 months into, the programme were analysed. Changes in intervention classes included increased use of open-ended questions by the teacher, increased participation of pupils in classroom discussion and development in critical reasoning. There were no changes in control classes. Variation in degree of

change between intervention schools was evident, and this may account for differences between intervention classes, rather than any wider differences between teachers.

Sustainability, replicability, cost effectiveness

This study provided evidence that it is possible to intervene effectively in the cognitive development of children of primary school age across a whole school district through a relatively light intervention of 1 hour per week with normal class sizes (up to 31) and pupil–teacher ratios. This is in contrast to some other thinking skills programmes – cf. the 5 hours each week of pencil and paper tasks and one-to-one mediation required in the Instrumental Enrichment programme (Feuerstein *et al.*, 1980). This study also demonstrated that it was possible to deploy collaborative enquiry successfully within the normal constraints of local authority funding and staff development time. There are important implications for sustainability and replicability here.

The cost effectiveness of the intervention was calculated by adding the intervention costs during the period of the evaluation, subtracting evaluation costs (because the evaluation was not a necessary requirement for such an initiative to function), and relating this to the measured cognitive outcomes. The costs included 10–12 hours of professional development per teacher, curricular materials for teachers and a supply teacher to cover the part-time involvement of the professional development leader. The total costs were divided by the number of teachers involved and then by the average number of pupils in each class. These calculations indicated an intervention cost of GB pounds £233 (US $410, Euro €345) per teacher or £9 ($16, €13) per pupil (further details from the authors on request). The teachers continued to have access to follow-up support and advice after the period of the evaluation. They were thus in a position to continue to develop their skills and develop new communities of enquiry with fresh classes. The cost per pupil for those classes was significantly less (£5.50, $10, €8 per pupil) than for the first group of pupils.

Given the evidence of programme effectiveness in terms of significant and generalized cognitive gains, cost effectiveness was considered to be high. However, while the evidence that middle and lower quartiles of pre-test ability benefit is encouraging, a stronger focus might be needed on adapting the programme to ensure that the top quartile benefit.

Theoretical modelling

These results suggest that the Philosophy for Children method can indeed increase children's 'intelligence', supporting the 'incremental' theory of intelligence. Beyond this, it appears that the generative, extending, relational and strategic aspects of P4C yield generalized cognitive effects – in non-verbal and quantitative as well as verbal aspects of intelligence. This may be seen as supporting the unitary 'g' factor notion of intelligence, and contradicting the proposal of multiple intelligences. However, such generalization may not be automatic, and should not be anticipated in other interventions lacking the elements fostering generalization which characterize P4C.

These findings are important in terms of potential impact of P4C upon subsequent national examination performance, which might determine a pupil's future opportunities to put their thinking skills to more creative use. Theoretical modelling suggested P4C incorporated many features likely to enhance generalization, and the data were in accord with this analysis. Such theoretical modelling might be useful prior to

similar studies of the generalization of other forms of thinking skill intervention. It might also help practitioners when designing adaptations of existing methods for higher generalization.

Future research

The finding that middle and lower quartiles of pre-test ability benefit most is encouraging, but more work is needed by both researchers and practitioners on how to ensure that the highest quartile of pre-test ability benefits equally. The influence of socio-economic factors on pupil outcomes (if any) remains unclear. There is some anecdotal evidence (e.g. Lake, 2000) that Philosophy for Children may be particularly helpful to children from more socially disadvantaged backgrounds whose language skills are less developed. It would be useful to investigate in more detail the relationship between the outcomes of philosophical enquiry and social and language factors. More detailed process investigation is also necessary more thoroughly to determine key elements of teacher behaviour in the intervention, and whether there are differences in the way that boys and girls respond to collaborative classroom processes.

An important issue is the duration of cognitive gains found in this study. Feuerstein (2004) has claimed that the cognitive gains arising from the Instrumental Enrichment programme are not only sustainable, but that the gains increase over time. Participants in the present study are being followed-up 2 years later, after they have transferred from primary to high school. The local authority is also initiating a programme of collaborative philosophical enquiry across high schools. Tracking the cohort will lead to comparison of pupils with no further experience of collaborative enquiry in high school with a group of pupils who continued to be involved in regular enquiry. This investigation is particularly relevant to concerns about a drop in the performance of pupils following their transition from primary to high school (e.g. Galton, Gray, & Rudduck, 1999).

Conclusions

This study found that, compared with non-intervention controls, weekly philosophical collaborative enquiry intervention over time led to:

(1) Significantly larger gains in measured overall cognitive ability;
(2) Significant gains in measured verbal cognitive ability, and also in non-verbal and quantitative reasoning ability;
(3) Gains in measured cognitive ability largely irrespective of school/class and pupil gender;
(4) Gains for all quartiles of pre-test ability, middle quartiles showing the biggest, the upper quartile the smallest;
(5) Effect sizes in measured cognitive ability from a large-scale field trial larger than those found in research studies in the literature.

Thus almost all the research questions were answered in the affirmative. P4C did appear to have significant positive effects and gains did generalize, as could be predicted from features of the method. Gains were largely consistent across participating schools. Encouraging gains were evident for middle- and for lower-achieving pupils. Boys showed somewhat higher gains than girls (although caution is needed here given the

lack of statistical significance and gender disproportionalities in the groups), interesting at a time when concerns have been raised about the extent to which girls academically outperform boys (e.g. National Foundation for Educational Research, 1999). Implementation costs for the intervention were very modest, indicating high cost effectiveness and potential for replicability and sustainability. There are implications for practice and policy, but also for the nature and completeness of future research on P4C and other thinking skills interventions.

While this study suggests that children benefit from collaborative enquiry, the positive outcomes also suggest a need to provide more opportunities to enable teachers to develop the relevant skills and dispositions in both initial teacher training and continuing professional development. Such experiences might usefully aim to provide teachers with an opportunity to develop their own critical thinking skills and the necessary confidence to use enquiry in the classroom.

Acknowledgements

This research was funded by Clackmannanshire Education Authority, Scotland.

References

Adey, P. (2001). Cognitive acceleration: Thinking as intelligence. *Teaching Thinking*, *5*, 38–41.

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

Burden, R., & Nichols, L. (2000). Evaluating the process of introducing a thinking skills programme into the secondary school curriculum. *Research Papers in Education*, *15*(3), 293–306.

Campione, J. C., Shapiro, A. M., & Brown, A. L. (1995). Forms of transfer in a community of learners: Flexible learning and understanding. In A. McKeough, J. Lupart & A. Marini (Eds.), *Teaching for transfer: Fostering generalization in learning* (pp. 35–68). Mahwah, NJ: Erlbaum.

Carnell, E., & Lodge, C. (2002). *Supporting effective learning*. London: Paul Chapman Publishing.

Cleghorn, P. (2002). *Thinking through philosophy*. Blackburn: Educational Printing Services.

Cobb, P., & Bowers, J. (1999). Cognitive and situated learning perspectives in theory and practice. *Educational Researcher*, *28*(2), 4–15.

Cotton, K. (2002). *Teaching thinking skills. School research series (SIRS)*. Portland, OR: Northwest Regional Educational Laboratory. Retrieved December 1, 2003, from www.nwrel.org/scpd/sirs/6/cu11.html.

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

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

Feuerstein, R. (2004, August). *The theory, the research and the techniques in action*. Presentation at Unlocked Potential Conference, Glasgow Caledonian University, Scotland.

Feuerstein, R., Rand, Y., Hoffman, M., & Miller, M. (1980). *Instrumental enrichment: An intervention program for cognitive modifiability*. Baltimore, MD: University Park Press.

Fisher, R. (1999). *Teaching thinking: Philosophical enquiry in the classroom*. London: Cassell.

Galton, M., Gray, J., & Rudduck, J. (1999). *The impact of school transitions and transfers on pupil progress and attainment*. London: Department for Education and Employment.

Gardner, H. (1999). *Intelligence reframed*. New York: Basic Books.

Goleman, D. (2005). *Emotional intelligence* (10th anniversary edition). New York: Bantam Books.

Haynes, J. (2001). *Children as philosophers*. London: Routledge Falmer.

Hernstein, R., & Murray, C. (1994). *The bell curve*. New York: Free Press.

Lake, M. (2000). A disturbing power to predict. *Teaching Thinking*, *1*(2), 20–25.

Lipman, M. (1981). Philosophy for children. In A. L. Costa (Ed.), *Developing minds: Programs for teaching thinking* (Vol. 2, pp. 35–38). Alexandria, VA: Association for Supervision and Curricular Development.

Lipman, M. (1991). *Thinking in education*. Cambridge: Cambridge University Press.

Lipman, M. (2003). *Thinking in education* (2nd ed.). Cambridge & New York: Cambridge University Press.

Lipman, M., & Sharp, A. M. (1978). *Growing up with philosophy*. Philadelphia, PA: Temple University Press.

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

Locke, A., Ginsberg, J., & Peers, I. (2002). Development and disadvantage: Implications for the early years and beyond. *International Journal of Language and Communication Disorders*, *37*(1), 3–15.

Lohman, D. F., Thorndike, R. L., & Hagen, E. P. (1993). *Cognitive abilities test*. Windsor: NFER-Nelson.

Matthews, G., Zeidner, M., & Roberts, R. D. (2002). *Emotional intelligence: Science and myth*. Cambridge, MA: MIT Press.

McGuinness, C. (1999). *From thinking skills to thinking classrooms: A review and evaluation of approaches for developing pupil's thinking* (Research Report RR115). London: Department for Education and Employment.

Mercer, N. (2000). *Word and minds: How we use language to think together*. London: Routledge.

Moseley, D., Baumfield, V., Higgins, S., Lin, M., Miller, J., Newton, D., Robson, S., *et al.* (2004). *Thinking skill frameworks for post-16 learners: An evaluation*. Guildford: Learning and Skills Development Agency.

National Foundation for Educational Research. (1999). *Boys achievement, progress, motivation and participation*. Windsor: NFER-Nelson.

Perkins, D. N., & Salomon, G. (1994). Transfer of learning. In T. Husen & T. Postlethwaite (Eds.), *The international encyclopaedia of education* (2nd ed., Vol. 11, pp. 6452–6457). Oxford: Elsevier/Pergamon.

Piaget, J. (1932). *The moral development of the child*. London: Routledge & Kegan Paul.

Shayer, M., & Adey, P. (2002). *Learning intelligence: Cognitive acceleration across the curriculum from 5 to 15 years*. Maidenhead: Open University Press.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Smith, P., Fernandes, C., & Strand, S. (2001). *Cognitive abilities test 3. Technical manual and pupil book (Levels B and C)*. Windsor: NFER-Nelson.

Sternberg, R., & Bhana, K. (1996). Synthesis of research on the effectiveness of intellectual skills programs: Snake oil remedies or miracle cures? *Educational Leadership*, *44*(2), 60–67.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J. (1997). *Successful intelligence*. New York: Plume.

Sternberg, R. J., & Jensen, A. R. (1992). Taking sides: Clashing views on controversial issues. In B. Slife & J. Rubenstein (Eds.), (pp. 144–165). Guildford, CT: Dushkin Publishing Group.

Thinking Skills Review Group. (2004). *Thinking skills approaches to effective teaching and learning: What is the evidence for impact on learners?* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Topping, K. J., & Ehly, S. W. (2001). Peer assisted learning. *Journal of Educational and Psychological Consultation*, *12*(2), 113–132.

Topping, K. J., & Trickey, S. (in press). Impact of philosophical enquiry on school students' interactive behaviour. *International Journal of Thinking Skills and Creativity*.

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

Watkin, C. (2001). Learning about learning enhances performance. *National School Improvement Network (NSIN) Bulletin*, *13*, 1–7.

Wegerif, R., Mercer, N., & Dawes, L. (1999). From social interaction to individual reasoning: An empirical investigation of a possible socio-cultural model of cognitive development. *Learning and Instruction*, *9*(5), 493–516.

Wenger, E., Pea, R., Brown, J. S., & Heath, C. (1999). *Communities of practice: Learning, meaning, and identity*. Cambridge & New York: Cambridge University Press.

Wilson, V. (2000). *Can thinking skills be taught?* Edinburgh, Scotland: Scottish Executive Education Department.

Received 5 October 2005; revised version received 8 February 2006