Evaluating Philosophy for Children: A Meta-Analysis

Félix García-Moriyón*; Irene Rebollo+£; Roberto Colom+

Affiliation

* Facultad de Formación del Profesorado
Universidad Autónoma de Madrid
Madrid (Spain)

+ Facultad de Psicología
Universidad Autónoma de Madrid
Madrid (Spain)

£ Biological Psychology Department
Vrije Universiteit
Amsterdam (The Netherlands)

Corresponding author

Irene Rebollo
Dept. of Biological Psychology, VU
Van der Boechorststraat 1
NL-1081 BT Amsterdam, The Netherlands
Tel: +31 20 4448981 fax: +31 20 4448832
E-mail address: i.rebollo@psy.vu.nl

Abstract

Philosophy for Children (P4C) is a learning-to-think program designed to foster high order reasoning skills and administered across the world. However, there is a lack of evidence concerning its efficacy to increase basic cognitive abilities. Only eighteen datasets among the more than 100 found were suitable to be submitted to meta-analysis, because most of them did not pass the standardized criteria. The results have several points of interest. First, P4C has a positive effect. The average difference between the treated (experimental) and untreated (control) groups is equivalent to half a standard deviation (d = .58). Second, researchers should use appropriate designs: only 15% of the revised articles passed minimum criteria. The

recommended design tests one experimental and one control group, before and after the program implementation. Third, researchers must provide necessary data: means, standard deviations, and number of participants, both for the pre-test and post-test measurements. Fourth, the dependent measure must be taken from available standardized measures of cognitive ability. Finally, the analyzed studies reveal that P4C is mostly administered only along a school year. We highly recommend the administration of P4C across school years in order to both increase the observed gains and make them enduring.

Introduction

Philosophy for Children (P4C) is a program aimed to develop and foster high order thinking skills. It was implemented for the first time at the United States in 1970. The creator of the program, Matthew Lipman, realized that many of his students lacked the basic critical thinking skills needed to cope with social and political problems, even with academic assignments. He started developing his curriculum based on two strong assumptions: according to Piaget's stages theory, children 12 year olds showed a growing ability to formal and abstract thinking; the Western philosophical tradition had attached too much importance to the *art* of good reasoning. The backbone of the new educational program was putting together those two simple and basic findings: he developed a curriculum where philosophical discussions about philosophical topics were the academic stuff that helped children to foster their thinking skills (Lipman, Sharp, & Oscanyan, 1975b).

Lipman's approach to education was very much influenced by Dewey's philosophy and educational proposals, such as those he set out in *Democracy and education* (Dewey, 1966) and put into practice at some schools in the States. Lipman followed Dewey's ideas about the practical aspects of teaching, such as curriculum development, educational methodologies, and the role of thinking and values in school system. Moreover, he accepted the particular emphasis Dewey himself placed on the strong relationship between education and the commitment to democracy as a specific social and political way of life. The influence of other American philosophers, such as Peirce and Mead was also very important. Peirce (Buchler, 1955) offered Lipman a specific understanding of the need of conceptual clarification and good thinking, and the importance of building a community of inquiry to foster both cognitive skills. From Mead (Mead, 1974), he took the idea of the social dimension of human personality that is one of the cornerstones of the community of inquiry approach to education. Another very important

influence comes from Wittgenstein (Wittgenstein, 1953), mainly from the second period of his philosophy when he pays attention to the analysis and clarification of everyday language. Most of the time, the discussions in the classroom doing philosophy with children focus on children's language and the basic philosophical concepts such as "time", "truth", "reality", beauty", and so forth.

Lipman's philosophical background is not limited to the American, or English, tradition. During his one year stay in France after the Second World War, he became familiar with the continental philosophical tradition. The ideas of Ricoeur, Balibar, and other French and European philosophers added new ideas, topics and procedures that he included in his educational project (García-Moriyón, 2002). This wider approach to philosophy was reinforced with the cooperation of her colleague in the development of the program, Ann Sharp, a woman familiar with Nietzsche's and Simone Weil's ideas. Hermeneutics, phenomenology, and existentialism are imbedded in the Philosophy for Children curriculum.

The curriculum was initially intended for 11 to 12 year olds. Later on, after Lipman rejection of Piaget's approach to cognitive development (Lipman, 1999), new materials were designed to foster basic thinking skills in elementary school and kindergarten. The advent of Philosophy for Children coincided with the recognition that emerged in the third quarter of the 20th century that children were capable of thinking critically and creatively (Matthews, 1980), and that a major aim of education should be to help children become more reasonable; Piaget's stages theory did not fit the new findings from psychological and educational research. More important than just to determine the developmental stage of children in order to use the appropriate materials to work in the classroom, was to discover what children would be able to do if the teachers offered them activities that actually challenged their cognitive and affective skills. As Bruner suggested (Bruner, 1977), your educational style depends on your psychological background, and it makes a difference if you move first from Skinner to Piaget, then to Vigotsky's ideas.

As the IAPC staff suggests "the last thirty years' experience in doing philosophy with children and adolescents has shown us that they are not only capable of doing philosophy but need and appreciate it for the same reasons that adults do. Children think constantly, and reflect on their thoughts. They acquire knowledge and try to use what they know. And they want their experience to be meaningful: to be valuable, interesting, just, and beautiful. Philosophy offers children the chance to explore ordinary but puzzling concepts, to improve their thinking, to make more sense of their world and to discover for themselves what is to be valued and cherished in that world" (IAPC Webpage)

The curriculum spanned from age 4 to 18: (a) From 4 to 12, to master the basic thinking skills, and (b) from 12 on to apply those skills to ethical, esthetical, and social topics. The curriculum is designed to get students involved in the exploration of the philosophical dimensions of their experience, with particular attention to logical, ethical, and aesthetic dimensions. These three areas of philosophy are related with critical (logic), caring (ethics) and creative (aesthetic) thinking, that is, with high order thinking. Lipman's curriculum consists of novels for students and manuals for teachers. Each novel is about 80 pages in length and is written in informal language, without technical terminology. Each manual is about 400 pages in length and contains conceptual explanations for teachers as well as discussion exercises and activities that can be used to supplement the students' inquiry. These manuals are indispensable for conducting dialogical inquiry (Accorinti, 1999).

Although P4C was first developed focusing on the cognitive dimensions of human mind, it broadened its goals and stressed the necessity of fostering affective and social skills (Lipman, 1999). New novels and manuals were designed to meet specific requirements of children's personal growth. First, Lipman moved to moral education, a serious problem in modern societies, offering an original approach to that area, one that pays attention to the moral growth of children (García-Moriyón, 1998). Children's experience is replete with ethical concerns and issues, and through the omnipresent media, children today are exposed to ideas and images which not so long ago would have been reserved for adults. Then, like adults, children often perceive the world as a jumble of alternative possibilities. Rather than dictate a set of prescribed values to children, Philosophy for Children seeks to help them strengthen their own capacity to appraise and respond to these puzzling alternatives; to self-correct their habits of thought, feeling and action through sustained ethical inquiry. Moreover, Philosophy for Children's egalitarian nature, commitment to varying viewpoints and insistence on the inherent value of all participants helps foster empathy and pro-social behavior as an essential basis for values in education (Sharp & Splitter, 1995).

The curriculum moved then to social problems, and a new novel and manual was designed for teenagers. Later on, Lipman and his collaborators went back to the firsts years of children: the new materials spread from age 5 to 11, with discussions about the experience, language, personal identity, ecology, ethics, and almost every other subject that call children's, and adults', attention (Lipman, 1988). After many years of sustained work, Philosophy for Children can offer teachers and schools a complete curriculum from 5 to 18 years; this is a coherent consequence of the basic assumptions of the program: if we want to foster children's

high order thinking, we have to do philosophy with then no just for a short period of their school time, but throughout all their school time, from kindergarten to high school (Garza Camino, 1994).

Lipman's approach got very soon an international recognition, and his curriculum spread all over the world, with translations into several languages and implementation in thousands of schools at many countries (I.A.P.C., 1982; I.A.P.C., 1991).

From the beginning, there has been a deep interest in the impact of the implementation of the program on the students (Weinstein, 1989). Lipman himself and his staff in the Institute for the Advancement of Philosophy for Children, established a close cooperation with the New Jersey Department of Education and its Educational Testing Service. The first step was to design a test to evaluate thinking skills, the New Jersey Test of Reasoning Skills (NJTRS,(Ellen, 1992)), focusing on the reasoning skills such as they appeared in the curriculum for children 11 years old. Then, they conducted a first field experiment in two schools at Montclair district (Lipman & Bierman, 1975a) and, two years later, a wider study involving thousands of children in different schools in New Jersey (Shipman, 1983).

Since then, several investigations on the effectiveness of the program have been conducted in many countries. A great part of that research uses a quantitative methodology with the NJTRS as the main evaluation tool. During the last years, some people have moved to more qualitative methods (Daniel, 2002; Pálsson, 1996; Santi, 1993), influenced by new trends in psychological and educational evaluation. On the other hand, though the program is also intended to modify other affective skills and personality traits, there are not many studies evaluating skills other than cognitive.

At present, most studies support the evidence of a positive impact on children's reasoning skills (I.A.P.C., 1982; I.A.P.C., 1991). However, the discussions concerning the program evaluation have always been controversial in the international and national conferences. The main topics under discussion are: a) the possibility of an evaluation of the skills fostered by the program; b) the skills fostered by the program that should be evaluated; c) the adequacy of quantitative or qualitative methods for a valid evaluation of those skills; and, d) the possible implications of the findings resulting from the program evaluation.

Despite all the work that has been done to evaluate the impact of P4C *there is not a clear conclusion yet*. The IAPC (Institute for the Advancement of Philosophy for Children) published two reviews of a selection of findings (I.A.P.C., 1982; I.A.P.C., 1991), and Lipman offered a summary of those findings in his first book (Lipman et al., 1975b). However, most of those papers have been criticized for being mainly intended to advertise or convince of the goodness of the program and

its efficacy, instead of critically inquiring about its actual impact (Morehouse, 1995; Sigurdadottir, 2002; Slade, 1992).

If one tries to investigate what has been done, the first problem encountered is t*he wide variety of approaches and designs used*. Some scholars just offer a short and simple description of their subjective positive (or negative) feelings after doing philosophy with children in their classrooms (Browning, 1988; Kyle, 1987; Schleifer & Poirier, 1996). Other scholars are very committed to a rigorous analysis and description of the categories that must be applied in a qualitative research(Daniel, 2002; Echeverría, 2003). Since the middle of the 80's, many scholars abandoned the quantitative methodologies and moved to qualitative designs, but the lack of a specific and shared methodology makes it difficult to get a clear understanding of the findings of their evaluations. Even if one focuses on the evaluations that use a classical quantitative methodology, several different designs are used, and it is rather difficult to reach a straight conclusion.

The second problem to face is the poor research reports. Most reports lack a complete and clear description of the employed methodology and the obtained results. Some reports give an incomplete amount of information, some do not present any data at all, and the majority does not follow the basic rules established by the scientific community for the presentation of a research report. This "state of affairs" could be explained by the philosophical background of the authors, and their scarce experience with the methodological requirements of educational and psychological research.

The third problem encountered is sample size. One can find well designed studies, with nicely presented results but, due to heterogeneous circumstances in the school or groups, the researcher works with a very small group of children(Charlann, 1979; Lipman et al., 1975a; Slade, 1988; Strohecker, 1986). It is risky to generalize results reached with 7, 25, or even 35 children.

Last but not least, most reports do not include any statistical analysis, or what they do is poor, insufficient, or inappropriate.

Summing up, there is a great amount of research on the implementation of the program. The results tend to offer some support for the positive effect of P4C, although there are also some evaluations that yield more skeptical results (García-Moriyón, Colom, Lora, Rivas, & Traver, 2002; Meyer, 1988). It is difficult to reach a clear conclusion about the full implications of the positive impact or about its long lasting effect. We actually do not know if the program is working or what is the scope of its impact. The disagreement about the appropriate method of evaluation, or at least the variety of approaches, is a serious obstacle to compare and accumulate the evidence of

the past 30 years. A similar problem comes from the difficulty to reach an accepted definition of reasoning skill.

The interest in the rigor of empirical research has always been active among scholars within the area of P4C. Some papers have offered important contributions to clarify the theoretical and methodological problems of the evaluation (Chervin & Kyle, 1993; Henderson, 1988; Santi, 1993). Two papers go further and offer a review of the research that has been done (Morehouse, 1995; Sigurdadottir, 2002), and they make very interesting suggestions moving to more rigorous and well designed evaluations.

A meta-analysis such the one we are presenting in this paper can offer a better understanding of the effect of the program in boys and girls' cognitive development. The meta-analytic techniques are intended to revise the evidence in a field when it is difficult to reach a straight conclusion with a narrative review of the literature. Using quantitative and objective criteria, we will be able to answer several important questions:

- Is there a relationship between the program application and the factors that it is intended to influence (reasoning skills)?
- What is the size of the observed relationship?
- Are the results obtained across studies homogeneous?
- If they are not,
- Which characteristics of the studies could explain the variability of the results?

The answer to these questions presumably will help to define an agenda for future research in Philosophy for Children.

Method

Selection of Studies and Inclusion Criteria

Four approaches were used to locate the sample of studies:

- Computer searches of both PSYCINFO and ERIC databases were conducted using as key term 'Philosophy for Children', without any field specification. This search, in May 2002, yielded 116 publications.
- The main journals of the program: *Thinking, Analytic Teaching, Critical and Creative Thinking, and Aprender a Pensar,* were revised looking for papers fitting the inclusion criteria.

- Unpublished reports were requested through two mailing lists: P4C-list@belnet.be (English speaking) and filoninos@listserv.rediris.es (Spanish speaking).
- A general request for unpublished or published reports was made at the NAACI Conference, in June 2002. Two papers were received from this request.

Most of the collected papers were excluded from consideration because they did not fit the inclusion criteria:

(1) To test the effectiveness of P4C to improve reasoning skills. Papers which included the keywords, but were not related to the program, or those with only theoretical aims, were excluded.

(2) To measure reasoning skills or mental abilities as the dependent variable. Papers using personality variables or not using any measure at all were excluded.

(3) To include enough statistical information to calculate the effect size associated to the effectiveness of the program to improve reasoning skills. The report had to include the sample sizes, means, and standard deviations, or the value of the *t*, or F tests, or the exact *p* value obtained from the previous tests. Studies which did not include any of those data, or only included graphs, frequencies, or verbal reports of the significance of the tests, were excluded.

Most of the collected papers were focused on the evaluation of reasoning skills (54), only two of them tested affective and cooperative skills isolated, and some of them (15) evaluated both cognitive and affective characteristics. Finally, twenty papers were found only with theoretical aims without any data.

Coding Procedure

Sixteen publications fitted the inclusion criteria. More than one study per publication was selected only if the samples among studies were independent. Two of the considered papers fitted this criterion (Martin & Weistein, 1985; Slade, 1989). Otherwise, only one study by publication was analysed. *Eighteen studies were submitted to meta-analysis*.

Each one of the 18 selected documents was independently coded by two expert P4C teachers. Initially, three kinds of moderator variables were coded (Lipsey, 1994):

Substantive:

- Measure: the instrument used to measure reasoning skills or mental abilities. Firstly, the specific name of the instrument was coded. Afterwards, this variable was recoded into two categories: (1) New Jersey Test of Reasoning Skills or Q4, and (2) others. The NJTRS was developed to measure the factors that the program is intended to change (the Q4 is

an earlier version of the same test). If the program has real effects on reasoning abilities, those must be detected by measures not directly related to P4C, but also measuring reasoning skills. All the studies provided this information and thus, this variable was analysed as a possible moderator under the hypothesis that measures external to the program would lead to smaller effect sizes.

- Mean age of the students: when the control and treatment group differed in age, the mean between them was computed. Some papers reported the academic grade from which the mean age could be inferred. The mean age was 11.54 (SD = 1.97; range = 8_15.65). There was not enough variability to differentiate between children and adolescents and thus, this variable was not analysed as a moderator.
- Book: The book used during the application of the program was coded: Harry (Lipman, 1982), Pixie (Lipman, 1981b), Lisa (Lipman, 1981a) or Mark (Lipman, 1980). Most of the studies followed Harry, developed to be used with children between 10 and 12 year olds. Thus, this variable was not considered as a moderator during the analyses.
- Teacher training: the amount of training and experience of the teacher who applied P4C could influence its effectiveness. Sadly most of the papers did not include this information and, when it was included, they always were considered experts. Thus, this variable was not analysed as a moderator.
- Administration time: given that the program of P4C is thought to be a longitudinal treatment, it is reasonable to think that the longer the application, the larger the effect. When the exact number of months was not reported, one academic year was coded as 9 months, and one term as 3 months. The mean duration of application was 7.33 months (SD = 3.77; range = 2_18). There was not enough variability to consider time of application as a moderator variable, considering that there were 6 studies with 9 months, and 4 studies with 8 months.

(1) Methodological:

- Research Design: there are two kinds of research designs that can be used in order to test the effectiveness of a treatment or program: Independent Groups and Repeated Measures. In the Independent groups design one group receives the treatment and the other group serves as control (untreated). Then the difference between the groups on the outcome measure is used as an estimate of the program effect. In the repeated measures design a single group is used, and each individual is measured before and after the treatment (pre and post test). Then the difference between the individual scores before and after the treatment is used as an estimate of the program effect. The independent

groups post-test design gives a biased estimation of the effect of the program, given that the difference between the groups already present before the program implementation is unknown. The single-group pretest-postest design is also biased because, without a control group, we do not know if the changes on the outcome measure are due to maturation or time. The most reliable and less biased design is a combination of the former two, testing an experimental and a control group before and after the program application. That way, the effects of maturation and prior differences between the groups are controlled. The studies submitted to meta-analysis were classified into these three categories to analyse the moderator effect of the applied research design. The hypothesis tested is directly related to the sources of bias present in each design. The independent groups + pretest-postest design will lead to the lowest effect sizes.

(2) Extrinsic:

- Year of publication: This factor was analysed as a moderator under the hypothesis that former studies will lead to larger estimations of the program effect, while more modern studies will give lower estimations. This hypothesis is related, not only to the willingness of the researchers to find nice effects when a project is starting, but mainly to the sophistication of the research designs and statistical techniques applied as a product of time and experience.
- Source of publication: This factor was considered under the suspicion that publications directly related to the program could bias the acceptance of papers towards those supporting the effectiveness of the program. Then external journals or editorials would tend to publish papers with lower effect sizes. We did not find enough variability to test this hypothesis: 13 out of 18 studies belonged to journals related to the program (Thinking and Analytic Teaching). Among the 5 other studies only one was from a peer reviewed publication (*Psichotema*).

Table 1 contains the information of the 3 variables finally considered for analyses (measure, design, and year) on the eighteen selected studies (Allen, 1988a; Allen, 1988b; Bierman, 1976; Camhy & Iberer, 1988; Cummings, 1979; García-Moriyón, Colom, Lora, Rivas, & Traver, 2000; García-Moriyón et al., 2002; García-Moriyón, Moreno, Pascual Díez, & Traver, 1988; Iorio, Weinstein, & Martin, 1984; Karras, 1979; Lipman & Bierman, 1976; Martin et al., 1985; Pálsson, 1996; Reed & Allen, 1982; Slade, 1989; Sprod, 1997).

Table 1. Moderator variables coding

REFERENCE	MEASURE	DESIGN	YEAR
(Allen, 1988a)	NJTRS o Q4	Independend groups pretest-postest	1988
(Bierman, 1976)	Others	Independent groups posttest	1976
(Camhy et al., 1988)	NJTRS or Q4	Single group pretest-postest	1988
(García-Moriyón et al., 2002)	Others	Independend groups pretest-postest	2002
(Cummings, 1979)	Others	Independend groups pretest-postest	1979
(García-Moriyón et al., 1988)	Others	Independent groups posttest	1988
(García-Moriyón et al., 2000)	Others	Independend groups pretest-postest	2000
(Iorio et al., 1984)	NJTRS or Q4	Single group pretest-postest	1984
(Karras, 1979)	Others	Independend groups pretest-postest	1979
(Lipman et al., 1976)	Others	Independent groups posttest	1976
(Martin et al., 1985)(1)	NJTRS or Q4	Single group pretest-postest	1985
(Martin et al., 1985) (2)	NJTRS or Q4	Single group pretest-postest	1985
(Martin et al., 1985) (3)	NJTRS or Q4	Single group pretest-postest	1985
(Pálsson, 1996)	NJTRS or Q4	Single group pretest-postest	1996
(Reed et al., 1982)	NJTRS or Q4	Independent groups posttest	1982
(Slade, 1989) (1)	NJTRS or Q4	Independend groups pretest-postest	1989
(Slade, 1989) (2)	NJTRS or Q4	Independend groups pretest-postest	1989
(Sprod, 1997)	Others	Independend groups pretest-postest	1997

Meta-Analytic Analyses

A meta-analysis or quantitative review answers three main questions:

(1) Which is the global effect size estimated from the selected publications? Is it statistically significant?

(2) Are the effect sizes obtained across studies homogeneous? If they are not, is there enough variability among them to look for possible explanations?

(3) Is there any model based on the characteristics of the studies that could explain the observed heterogeneity?

To answer the first question it is necessary to translate the data of the primary research into a common metric, that is, the effect size (*d*). Given that there are three kinds of research designs that can be used to test the effectiveness of a given program, the data obtained from each one is not directly comparable. Different estimates of the effect size were applied to each design according to Morris & DeShon (2002).

To estimate the mean effect size, each study's effect size must be weighted by its sampling variance. The sampling variance is also affected by the research design, and Morris & DeShon's (2002)advice is applied again.

Once the mean effect size is estimated, its heterogeneity must be tested by a significance test: the *Q* statistic. This statistic tests, against a chi-square distribution, the null hypothesis that there are not differences among the effect sizes of the primary studies and thus, all of them are estimations of the same parameter. If the Q statistic is statistically significant, it can be inferred that the variability across studies is not due to sampling or random effects, and different studies are estimating different parameters; some explanation must be found.

If the Q statistic yields a significant value, the third step is to find systematic sources of variation among effect sizes. Those possible sources are the so-called moderator variables. To estimate the effect of categorical moderators an ANOVA is computed. To estimate the effect of quantitative moderators, a regression analysis is applied.

Results

-analysis.

Table 2. Sample sizes (N), *Conversion to a common metric*

Table 2 shows the sample sizes and the Effect Sizes (*d)* estimated for each primary study selected for meta Effect Sizes (*d*) and sampling variances (w_i) of the primary studies

REFERENCE	N Control	N Experim.	*d*	95% Confidence Interval for the *d*		w_i
				Min	Max	
(Allen, 1988a)	22	23	.6829	.6306	.7352	10.08
(Bierman, 1976)	14	14	.3944	.3421	.4467	6.32
(Camhy et al., 1988)		69	.6380	.5857	.6903	55.24

(García-Moriyón et al., 2002)	58	75	-.2284	-.2808	-.1761	32.00
(Cummings, 1979)	15	14	.7768	.7244	.8291	6.17
García-Moriyón, F. 1988	150	139	.3367	.2844	.3890	70.63
(García-Moriyón et al., 2000)	59	56	.2115	.1591	.2638	28.06
(Iorio et al., 1984)		336	.7873	.7350	.8396	254.45
(Karras, 1979)	64	64	.5394	.4871	.5917	30.35
(Lipman et al., 1976)	20	20	.8580	.8057	.9103	8.60
(Martin et al., 1985)(1)		287	.6933	.6410	.7457	229.32
(Martin et al., 1985) (2)		428	.5000	.4477	.5523	378.38
(Martin et al., 1985) (3)		249	.5097	.4573	.5620	218.31
(Pálsson, 1996)		62	1.3187	1.2664	1.3710	31.37
(Reed et al., 1982)	35	51	.7106	.6582	.7629	19.05
(Slade, 1989) (1)	15	15	.3218	.2694	.3741	6.86
(Slade, 1989) (2)	10	10	.1708	.1184	.2231	4.42
(Sprod, 1997)	29	28	.5469	.4946	.5992	13.20

The effect size or *d* is the number of standard deviations that separates the compared scores. For the studies using a repeated measures design, *d* is the number of standard deviations that the participants have improved between the pre and post tests. For the studies using an independent groups design, *d* is the number of standard deviations that separate the scores of the control and the experimental group after the program implementation. A negative *d* means that the scores of the experimental group have worsened instead of improved.

All the effect sizes except one are positive. All of them are significantly different from zero, as their confidence intervals do not include a zero value. Apart from the negative one, the smallest value is .1708 and the largest one is 1.318. Thus, the P4C program has a positive effect on the student's reasoning skills.

Combination of the Effect sizes

The estimation of the mean effect size weighted by each study's sampling variance (also shown in Table 1) yielded a value of .5848 ($p < .001$; CI = .5325_.6372). What this means is that the implementation of P4C leads to an improvement of the student's reasoning skills of more

than half a standard deviation. It must be noted that the majority of the studies implemented the program for a period of one academic year.

Heterogeneity test

The total variability estimated through the Q statistic is equal to 66.369 ($p < .001$). What this means is that there is some variation across studies in the sample sizes that must be explained by something different from sampling or random effects.

Moderator analyses

Table 3 shows the results obtained from the Analysis of Variance for the moderator "measure".

Table 3. ANOVA for Measure

Measure	*d*	95% C.I.		Q_{wi}	p (of Q_w)
		Min.	Max.		
NJTRS or Q4	.6272	.5325	.6372	13.4801	.0193
Others	.3106	.1703	.4508	47.1712	.0000
TOTALS	Q_B = 17.073 (p<.001)				
	Q_w = 49.296 (p<.001)				

Note. Q_{wi} is the variability existent into each category of the moderator; Q_b is the variability explained by the model; Q_w is te sum of the categories' variabilities and thus the remaining heterogeneity unexplained by the model.

The results indicate that there is a significant effect of the instrument used to measure reasoning abilities. Those studies using instruments directly related to the program obtained larger effect sizes than those using measures external to the program. But it must be noted that the effect size estimated for studies using other measures (0.3106) is still significantly different from zero. The remaining heterogeneity unexplained by this moderator is still significant, which means that there must be other characteristics of the studies influencing the effect sizes. The type of instrument used explains 25% of the heterogeneity.

Table 4 shows the results obtained from the ANOVA for the moderator "design".

Table 4. ANOVA for Design

Design	*d*	95% C.I.		Q_{wi}	p (of Q_w)
		Min.	Max.		
Independent groups post-test	.4511	.2596	.6428	3.650	.3018
Single group pretest-postest	.6310	.5736	.6883	31.651	.0000
Independend groups pretest-postest	.2810	.1099	.4522	14.6092	.0413
TOTALS	Q_B = 16.459 (p<.001)				
	Q_w = 49.910 (p<.001)				

Note. Q_{wi} is the variability existent into each category of the moderator; Q_b is the variability explained by the model; Q_w is te sum of the categories' variabilities and thus the remaining heterogeneity unexplained by the model.

The results indicate that there is a significant design effect. The independent groups post-test design yields the larger estimations, while the independent groups with pre and post test design yields the smallest estimation. Given that the latter is the less biased and most reliable design, we should take the effect size associated with it as the best estimation of the population parameter. Although it is lower than the mean estimation, .2810, it is still significant.

Another interesting result is that, only within the second category, there is still significant variability unexplained by this moderator. The type of design used explains 75.2% of the variability observed across studies.

Finally, a regression analysis was performed to study the possible effect of year of publication. The variance explained by the model is 9.2861 (p < .001), and the regression coefficient was -0.019. The negative coefficient indicates that the more recent the research, the lower the effect size obtained. However, it must be noted that the year of publication is related to the type of design applied (eta = .461) so the more recent the study, the higher the probability to use a complete design (independent groups with pre and post test).

Discussion

There are several learning-to-think programs administered across the world (Neisser et al., 1996; Nickerson, Perkins, & Smith, 1985). Virtually all of them are intended to foster high order cognitive abilities, like deductive and inductive reasoning, language, or decision making (Feuerstein, Rand, Hoffman, & Miller, 1980; Herrnstein, Nickerson, Sanchez, & Sweets, 1986). Philosophy for Children is among those programs (Lipman, 1976).

The evidence concerning the positive impact of these programs is scarce and widespread. There are very little efforts to test that impact. However, these efforts should be seen as crucial, because applied educational psychologists must take some criteria to decide among possible alternative programs (Grotzer & Perkins, 2000).

The meta-analytic approach is especially useful to gain knowledge about the impact of these learning-to-think programs. Moreover, the analysis of the published studies against several criteria has the additional advantage of highlight the inappropriate practices doing research and reporting results. Thus, for instance, research designs are usually far from ideal to test the effectiveness of a given program, basic data are unreported, or sample sizes are very small.

For those reasons, we submitted to meta-analysis several reports testing the effectiveness of P4C. This meta-analysis was not previously attempted. The results observed have several points of interest.

Nevertheless, before discussing the specific meta-analytic results we found, some comments about the theoretical program underpinnings are especially germane. According to the assumptions of Lipman and his associates (Lipman, 1993), teachers have to put much more emphasis in the development of high order thinking skills, and that specific cognitive enrichment of children has to start as soon as possible. A second point in his critical analysis of education is that doing philosophy with children is a very suitable way of fostering those thinking skills. With the adequate curriculum, adapted to the children's interest and their level of personal growth, philosophy will empower children's cognitive dimension.

Scholars, both from the philosophical and psychological field received this proposal with a deep scepticism (Lipman & Sharp, 1978). Academic philosophers used to consider philosophy one too much abstract subject matter, far away from children's abilities and interest. Educational and developmental psychologists looked at Lipman's approach with suspicious, insofar as he challenged Piaget's widely accepted ideas about concrete operational stage of cognitive development (Norris & Ennis, 1989; Swartz & Perkins, 1990).

The present meta-analysis offers a valuable support for Philosophy for Children. Teachers did philosophy in their classes, following the methodological directions presented in the teachers'

manuals and in the theoretical studies. They provided philosophical discussion to their students and stressed the importance of a rigorous thinking that takes care of deductive and inductive tools of good thinking. The evaluation of the implementation of the program confirms the prediction stated by the theory: children do improve their cognitive skills, and these results verify the program effectiveness. Therefore, *children can do philosophy and this practice helps them to develop their high order thinking skills*.

Returning our attention to the most relevant meta-analytic findings, it is safe to state that P4C has a positive effect over the target thinking or reasoning skills. The average computed effect size is a noteworthy 0.58. This value implicates that the gap between the treated (experimental) and untreated (control) groups is equivalent to more than half a standard deviation.

For illustrative purposes, note that if the *z* score of the control group is 0 and the *z* score of the treated group is 0.5 (a difference equivalent to half a standard deviation), then the corresponding IQ scores will be 100 and 107, respectively. A *group difference* of 7 IQ points is usually considered as a great difference (Hemphill, 2003; Nickerson et al., 1985).

Thus, for instance, if we fix a cutting IQ point of 120, then there will be 20% of people from the treated group falling beyond that point, while there will be 9% of people from the untreated group falling beyond this same cutting point. This clearly makes a huge difference.

The result is especially impressive if we note that P4C was applied only along a school year in all the revised studies. It is well known that the program was designed to be applied across several years (Lipman, 1976). It would be predicted that if the program is implemented during several years, the effect will be both much higher and enduring. However, it must be recognised that we do not have evidence to support that statement.

However, the average effect size must be carefully interpreted because significant variability across studies was found. That result indicates that different studies are estimating different population parameters for the effect size.

The dependent measure applied in the study was one of the characteristics causing variations in the effect size. The tests employed to compare the treated and untreated groups do make a difference in their observed gap. The New Jersey Thinking Skills Test was designed as one P4C very proxy measure. Unsurprisingly, the greater effect sizes were observed for that test. This practice must be avoided as far as possible. The selected dependent measure must be taken from the available measures that assess reasoning ability. Researchers can administer tests like the Culture Fair Intelligence Test, the Progressive Matrices Test, the Differential Aptitude Battery (DAT), or the Primary Mental Abilities Battery (PMA). The composite measure that can be derived

from those batteries assesses the so-called general cognitive ability, very closely related to reasoning ability (Carrol, 1993; Jensen, 1998).

Nevertheless, it is noteworthy that even using a dependent measure like the Culture Fair Intelligence Test, P4C still has a remarkable positive effect. Although it must be recognised that the effect is lower than the observed for the New Jersey Test, it is still significant.

Furthermore, the employed research design also has an effect on the observed results. We can take advantage of this result to urge researchers to use a standard design from now on. It is highly desirable to share a common design, not only for comparative purposes.

Morris and DeShon (2002) article gives a clear picture of the preferred design for these studies testing the effectiveness of a learning-to-think program. The most reliable and less biased design tests an experimental (treated) and a control (untreated) group before (test) and after (retest) the program implementation. That way, the effects of maturation and previous differences between the groups are controlled. We highly recommend researchers to follow this basic design.

Moreover, researchers must provide the appropriate data: means, standard deviations, and number of participants, both for the pre-test and post-test measurements. These data are necessary in order to facilitate possible re-analyses like the performed in the present article.

In summary, the results of the reported meta-analysis reveal that P4C has a positive effect. The search for empirical reports allowed us to note that the research efforts are highly widespread. This is clearly undesirable to compare studies performed in different countries, by different researchers, and with different samples. We highly recommend researchers to follow the guidelines proposed in the present paper. Moreover, P4C should be implemented according to its design: not only for 1 school year, but intensively across several school years. This will promote a greater advantage of the treated groups than the one currently observed.

REFERENCES

Accorinti, S. (1999). *Introducción a la Filosofía para Niños [Introduction to Philosophy for Children]*. Buenos Aires: Manantial.

Allen, T. (1988a). I think, therefore I can: Attribution and Philosophy for Children. *Thinking, The Journal of Philosophy for Children, 8,* 14-18.

Allen, T. L. (1988b). Doing philosophy with children. *Thinking, The Journal of Philosophy for Children, 7,* 23-28.

Bierman, M. L. (1976). A pilot study in the teaching of logic research conclusions. *Metaphilosophy, 7,* 35-39.

Browning, b. (1988). Harry in three classes. *Analytic Teaching, 8,* 70-72.

Bruner, J. (1977). *The process of Education*. Cambridge: Harvard University Press.

Buchler, J. (1955). *Philosophical writings of Perice*. New York: Dover Publications.

Camhy, D. G. & Iberer, G. (1988). philosophy for children: a research project for further mental and personality development of primary and secondary school pupils. *Thinking, The Journal of Philosophy for Children, 7,* 18-25.

Carrol, J. B. (1993). *Human Cognitive abilities. A survey of factor analytic studies*. Cambridge: Cambridge University Press.

Charlann, S. (1979). Philosophy for students with learning disabilities. *Thinking, The Journal of Philosophy for Children, 1,* 21-34.

Chervin, M. I. & Kyle, J. A. (1993). Collaborative inquiry research into children's pilosophical reasoning. *Analytic Teaching, 13,* 11-32.

Cummings, N. P. (1979). Improving the logical skills of fifth graders. *Thinking, The Journal of Philosophy for Children, 1,* 90-92.

Daniel, M. F. (2002). The development of dialogical critical thinking.

Ref Type: Unpublished Work

Dewey, J. (1966). *Democracy and Education*. New York: The Free Press.

Echeverría, E. (2003). El aprendizaje y la utilización del pensamiento crítico. Una investigación etnográfica. *Aprender a Pensar, 5,* 60-69.

Ellen, A. S. (1992). Review of the New Jersey Test of Reasoning Skills. In J.J.Kramer & J. C. Conoley (Eds.), *Eleventh Mental Measurements Yearbook* (pp. 606-608). Lincoln, Nebraska: Buros Institute of Mental Measurement.

Feuerstein, R., Rand, Y., Hoffman, M. B., & Miller, R. (1980). *Instrumental enrichment*. Baltimore, MD: University Park Press.

García-Moriyón, F. (1998). *Crecimiento moral y Filosofía para Niños [Moral growth and Philosophy for Children]*. Bilbao: Desclée de Brouwer.

García-Moriyón, F. (2002). *Matthew Lipman: Filosofía y educación [Matthew Lipman: Philosophy and Education]*. Madrid: Ed. De la Torre.

García-Moriyón, F., Colom, R., Lora, S., Rivas, M., & Traver, V. (2000). Valoración de 'Filosofía para Niños': un programa de enseñar a pensar. *Psicothema, 12,* 207-211.

García-Moriyón, F., Colom, R., Lora, S., Rivas, M., & Traver, V. (2002). *La estimulación de la inteligencia*. Madrid: Ediciones de la Torre.

García-Moriyón, F., Moreno, A., Pascual Díez, F., & Traver, V. (1988). Evaluación de la aplicación del programa de Filosofía para Niños [Evaluation of the application of the program Philosophy for Children.

Ref Type: Unpublished Work

Garza Camino, M. T. (1994). *Educación y Democracia [Education and Democracy]*. Madrid: Visor.

Grotzer, T. A. & Perkins, D. N. (2000). Teaching Intelligence: A performance conception. In R.Stenberg (Ed.), *handbook of Intelligence* (pp. 492-515). Cambridge: Cambridge University Press.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficient. *American Psychologist, 58,* 78-80.

Henderson, A. (1988). Program evaluation issues and analytic teaching. *Analytic Teaching, 8,* 43-55.

Herrnstein, R., Nickerson, R., Sanchez, M., & Sweets, J. A. (1986). Teaching thinking skills. *American Psychologist, 41,* 1279-1289.

I.A.P.C. (1982). Philosophy for Children: Where we are now. *Thinking, The Journal of Philosophy for Children, Supplement 2*.

I.A.P.C. (1991). *Philosophy for Children. A report on Achievement* Montclair State University.

Iorio, J., Weinstein, M., & Martin, J. (1984). A review of district 24's philosophy for children program. *Thinking, The Journal of Philosophy for Children, 5,* 28-35.

Jensen, A. (1998). *The g factor*. London: Praeger.

Karras, R. W. (1979). Final evaluation of the pilot program in philosophical reasoning in lexington elementary schools 1978-1979. *Thinking, The Journal of Philosophy for Children, 1,* 26-32.

Kyle, J. (1987). Not a success story: Why P4C did not 'take' with gifted students in a summer school setting. *Analytic Teaching, 7*.

Lipman, M. (1976). Philosophy for Children. *Metaphilosophy, 7*.

Lipman, M. (1980). *Mark*. Montclair: Institute for the Advancement of Philosophy for Children: University Press of America.

Lipman, M. (1981a). *Lisa*. Montclair: Institute for the Advancement of Philosophy for Children: University Press of America.

Lipman, M. (1981b). *Pixie*. Montclair: Institute for the Advancement of Philosophy for Children: University Press of America.

Lipman, M. (1982). *Harry Stottlemeier's*. Montclair: Institute for the Advancement of Philosophy for Children: University Press of America.

Lipman, M. (1988). *Philosophy goes to school*. Philadelphia: Temple University Press.

Lipman, M. (1993). *Thinking Children and Education*. Dubuque: Kendall/Hunt Publishing.

Lipman, M. (1999). *Natasha*. Montclair: Teachers College Press.

Lipman, M. & Bierman, M. L. (1975a). Field experiment in Montclair. In M.Lipman, A. Sharp, & F. Oscanyan (Eds.), *Philosophy goes to school* (Philadelphia: Temple University Press.

Lipman, M. & Bierman, M. L. (1976). Philosophy for Children; Appendix B: Experimental research in philosophy for children. *Metaphilosophy,* 217-224.

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

Lipman, M., Sharp, A., & Oscanyan, F. (1975b). *Philosophy goes to school*. Philadelphia: Temple University Press.

Lipsey, M. W. (1994). Identifiying potentially interesting variables and analysis opportunities. In H.Cooper & L. V. Hedges (Eds.), *The Handbook of REsearch Synthesis* (New York: Russell Sage Foundation.

Martin, J. F. & Weistein, M. I. (1985). Thinking skills and philosophy for children: the Bethlehem Program, 1982-1983. *Analytic Teaching, 5,* 28-31.

Matthews, G. (1980). *Philosophy and the young child*. Cambridge Mass.: Harvard University Press.

Mead, G. H. (1974). *Mind, self & society*. Chicago: The University of Chicago Press.

Meyer, J. R. (1988). A quest of the possible? Evaluation of the immpact of the Pixie programme on 8-10 years old. *Analytic Teaching, 9,* 63-64.

Morehouse, R. (1995). Research in Philosophy for Children: an outline and an agenda. *Critical and Creative Thinking: the Australian Journal of Philosophy for Children, 3,* 74-82.

Morris, S. B. & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7,* 105-125.

Neisser, U., Boodoo, G., Bouchard, T., Boykin, A., Brody, N., Ceci, S. et al. (1996). Intelligence: knowns and unknowns. *American Psychologist, 51,* 77-101.

Nickerson, R., Perkins, D. N., & Smith, E. (1985). *The teaching of thinking*. Hillsdale, NJ: LEA.

Norris, S. P. & Ennis, R. H. (1989). *Evaluating critical thinking*. Pacific Grove: Critical Thinking Press & Software.

Pálsson, H. (1996). We think more than before about others and their opinions. *Thinking, The Journal of Philosophy for Children, 12,* 24-28.

Reed, R. & Allen, H. (1982). Analityc thinking for childre in Fort Worth Elementary schools. Initial evaluation report. *Analytic Teaching,* 5-12.

Santi, M. (1993). Philosophizing and learning to think: som proposals for a qualitative evaluation. *Thinking, The Journal of Philosophy for Children, 10*.

Schleifer, M. & Poirier, G. (1996). The effect of philosophical discussion in the classroom on respect for others and no-stereotypic attitudes. *Thinking, 12*.

Sharp, A. & Splitter, L. (1995). *Teaching for better thinking. The classroom community of inquiry*. Camberwell: ACER.

Shipman, V. C. (1983). Evaluation replication of the philosphy for children program-Final report. *Thinking, The Journal of Philosophy for Children, 5,* 45-57.

Sigurdadottir, B. (2002). Overarchin statement: research.

Ref Type: Unpublished Work

Slade, C. (1988). Logic in the classroom. *Thinking, The Journal of Philosophy for Children, 8*.

Slade, C. (1989). Logic in the classroom. *Thinking, The Journal of Philosophy for Children, 8,* 14-20.

Slade, C. (1992). Creative and critical thinking, an evaluation of Philosophy for Children. *Analytic Teaching, 13,* 25-36.

Sprod, T. (1997). Improving scientific reasoning through Philosophy for Children: an empirical study. *Thinking, The Journal of Philosophy for Children, 13,* 11-16.

Strohecker, M. (1986). Results of the 1983-84 Philosophy for Children experiment in Lynbrook. *Thinking, The Journal of Philosophy for Children, 6*.

Swartz, R. J. & Perkins, D. N. (1990). *Teaching Thinking: Issues & Approaches*. Pacific Grove: Critical Thinking Press & Software.

Weinstein, M. (1989). The Philosophy of Philosophy for Children. *Analytic Teaching, 10*.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell Publisher.