Meta-Analysis of the Effectiveness of Philosophy for Children Programs on Students' Cognitive Outcomes

Sijin Yan, Lynne Masel Walters, Zhuoying Wang, and Chia-Chiang Wang

ABSTRACT: *Philosophy for Children (P4C) is an educational program that aims at introducing philosophy into K-12 education. This meta-analysis examines the research on P4C, published from 2002 to 2016, regarding how it affects pre-collegiate students' cognitive outcomes. Ten studies (including two follow-up studies) with the total sample size of 1,509 students from second to twelfth grade are included in this meta-analysis. Results suggest that the extant empirical studies on P4C show an overall moderate positive effect (d=0.58) on students' cognitive learning outcomes and a significant positive effect on reasoning skill (d=1.06). Specifically, those studies conducted in non-Western countries have higher effect sizes than the Western ones. Moreover, studies with smaller sample sizes have higher effects sizes than those with larger sample sizes. This may be because P4C produces better outcomes in reasoning skills than general cognitive abilities and reading comprehension, and P4C could be more effective when practiced in small scales.*

Keywords: Meta-analysis; Philosophy for Children; Cognitive Outcomes; Reasoning Skills

Philosophy for Children (often abbreviated as P4C) is an educational program that provides students in K-12 settings opportunities to engage in communities of philosophical inquiry with the long-term aim of improving their cognitive abilities (Lam, 2012; Trickey & Topping, 2004). An increasing number of studies have documented the implementation of P4C and its impacts on students' cognitive outcomes (Abbasi & Ajam, 2016; Lam, 2012; Nia, 2014). This meta-analysis aims to examine what the cumulative research evidence suggests about how P4C affects students' cognitive abilities and whether characteristics of interventions, students, or outcome types influence the magnitudes of the effectiveness of the program.

Introduction of Philosophy for Children

Philosophy for Children is an educational program initiated by Matthew Lipman, Ann Sharp and their colleagues in the Institute for the Advancement of Philosophy for Children (IAPC) in the early 1970s (Brandt, 1988; Lam, 2012; Marashi, 2008; van der Straten Waillet, Roskam, & Possoz, 2015). Witnessing the weaknesses of college students' argumentative performance in public discourses and the tumultuous political environment during the contentious years of Vietnam War in the 1970s (Vansieleghem & Kennedy, 2012), Lipman argued that philosophy should no longer be confined to college and academic research. Children, he said, even in elementary grades, can begin a quest in philosophy to learn how to think and reason (Brandt, 1988).

Currently, with the help of numerous philosophers, educators and researchers, Philosophy for Children has become a global movement that has spread across 50 countries, and its material has been translated into 20 languages (Daniel & Auriac, 2011). It is now a fertile ground for educators to creatively practice different educational ideas (Meskin & Cook, 2012; Murris, 1992) to critically modify it for different students living in specific cultural, social and educational contexts in the world (Di Masi & Santi, 2016; Ghazinejad & Ruitenberg, 2014; Ndofirepi & Cross, 2015; Ndofirepi & Shanyanana, 2016), and most recently, to introduce it to very young children who are below the age of 6 (Giménez-Dasí, Quintanilla, & Daniel, 2013; Säre, Luik, & Tulviste, 2016).

As a whole, Philosophy for Children believes in engaging pre-collegiate students in doing philosophy by 1) removing the formidable terminologies and 2) using children's literature, pictures, films or other forms of stimuli to bring the philosophical discussion into class.

The central pedagogy of Philosophy for Children is community of inquiry. Its philosophical roots can be traced back to the American philosopher Charles Sanders Peirce. In his article *Some Consequences of Four Incapacities* (1868), he claimed that it is pernicious to make single individuals absolute judges of truth and what we need is a community of inquiry to 'grind off the arbitrary and the individualistic character of thought' (Peirce & Houser, 1998). Dewey fleshed out the Peircean theory of inquiry and incorporated it into his philosophy of education (Lipman, 2004), contending that a genuine community life in the classroom could let students cultivate the habit to meaningfully engage in a democratic life. Ann Sharp reconstructed and put forward the notion of community of inquiry as the guiding educational model for the philosophy for children movement (Gregory & Laverty, 2017).

Here is an illustration of a classic example of a Mendham P4C session (noting that there are many other important forms of P4C and the ways of doing community of inquiry are always amendable): in a classroom of philosophy for children, students with diverse backgrounds and lived experiences gather together in a circle to read a selected text aloud. Normally each student reads a part of the text and everyone has a turn, so that they could share meaning with each other, read aloud with expression and emotions, and learn to carefully listen to others. Then, teachers collect students' questions about issues they find puzzling, and write them down on the chalkboard/whiteboard for further discussion. After students choose a question as the target of discussion, they inquire with each other, which often involves making assertions that are supported with reasons, clarifying one's position, and providing (counter) examples. The goal of such inquiry is to "form a judgement about the matter that is reasonable, meaningful and practicable as they can manage (Gregory & Laverty, 2017)."

However, given the divided political and racial climates in the current era of globalization, a community may not always be a birthplace of respect, diversity and mutual learning; it can also generate conflicts, discontent, feeling of divergences, bullying and exclusion. Thus, the idea of building 'an intellectually safe community' where all participants can be challenged in their worldviews but at the same time feel supported and safe is a timely response to such concerns in P4C (Butnor, 2012).

Literature Review

The radical nature of P4C in transforming our vision of the function of philosophy from a sphere for intellectual elites to a place for human beings with diverse age and life experiences generates skepticism and debate. The first question is whether children are intellectually mature enough for philosophy (Daniel & Auriac, 2011). This debate can be traced back to Plato and Aristotle's views on the nature of children and their negative opinions on children's intellectual ability and appropriateness to philosophize (Kennedy, 2006). In modern educational theories, according to Piaget, children are not equipped with the ability to do abstract thinking (Piaget, 1931). However, Vygotsky's emphasis on the role of social interactions in cultivating children's intelligence potential provides strong support for the idea that age is not the determining factor of children's cognitive abilities (Roberts, 2016). Recently, cognitive scientists have shown that children have much higher level of cognitive abilities than Piaget estimated (Gopnik, 2009). Furthermore, philosopher Gareth Matthews stressed the freshness of children's ideas that he discovered through his philosophical discussion with young people, and thus rejected deep-rooted condescending attitudes behind this 'children are not capable of doing philosophy' argument (Matthews, 1980, 1994).

The second question concerns the evaluation of P4C, which is related to questions such as whether P4C is effective according to different metrics and if there are various ways this program can be implemented to benefit more students, especially those who are challenged and disadvantaged, at an affordable cost (Gorard, Siddiqui, & Huat See, 2015).

Since the 1970s, the outcomes measured in P4C research can be divided into two categories: (1) cognitive outcomes (2) socio-psychological outcomes related to attitudes toward academics, prosocial attitudes and behavior. Even though there are emerging studies that appraise the effectiveness of P4C in the socio-psychological field (Abbasi & Ajam, 2016; Dasí, Quintanilla, & Daniel, 2013; Scholes et al., 2016), the extant literature is still limited. Most studies focus on the goal of the Philosophy for Children program to provide a more formal training to develop students' cognitive outcomes. This includes direct assessment of students' reasoning skills and abilities, comprehensive skills, and academic performance (García-Moriyón, Rebollo, & Colom, 2005; Gorard et al., 2015; Gregory, 2011; Säre et al., 2016; Trickey & Topping, 2004) as indicators to students' learning progresses.

Previous Review of the Evaluation of P4C

P4C may have a positive effect on students' cognitive abilities. In 2004 and 2005, two systematic reviews (García-Moriyón et al., 2005; Trickey & Topping, 2004) were conducted to synthesize research on the effectiveness of P4C. First, the quantitative systematic analysis by Trickey and Topping (2004) investigated the influence of P4C on students in general, with the conclusion that P4C has a moderate positive effect on students' abilities with low variance. It collected eight controlled experiments regarding Philosophy for Children from 1970s to 2002.

Even though the relationship between the two has not been yet been accepted by researchers (Hidi, Renninger, & Krapp, 2004), they combined the cognitive outcomes and affective abilities without a theoretical foundation for doing so.

The second study is a meta-analysis conducted by García-Moriyón, Rebollo, and Colom (2005), in which they examined the relationship between P4C and reasoning skills as outlined in 18 studies published from 1976 to 2002, with the finding that P4C has a positive moderate influence on students' reasoning abilities. This meta-analysis included 18 experiments. The results showed significant differences among post-test experiments, single group studies with pre and post-test, and controlled experiments, in which the more rigorous controlled experiments tended to show lower effect sizes.

These two reviews (García-Moriyón et al., 2005; Trickey & Topping, 2004) provided significant contributions in understanding the impacts of implementing Philosophy for Children in K-12 education. However, a new meta-analysis is needed to address the following issues in the contemporary situation: First, after the publication of the two earlier meta-analyses, a larger collection of literature on the effects of P4C on cognitive outcomes has been generated with an increasing rigor of study designs, a larger number of participants, and follow-up studies (Fair et al., 2015a; Fair et al., 2015b; Topping & Trickey, 2007a, 2007b). Thus, researchers now have the opportunity to improve the rigor of a systematic analysis by only including studies with random controlled experiments or quasi-experiments and analyze these findings in detail through moderator analyses to find if the relationship between cognitive outcomes and P4C intervention is depended upon other variables such as duration of the program, sample size, et cetera. Second, since the P4C movement has spread worldwide and research was conducted on different continents (Lam, 2012; Marashi, 2008; Nia, 2014; Youssef, 2014), a meta-analysis at this stage can involve an exhaustive search globally in English and capture the multiplicity of P4C practices worldwide. Thus, the present study aims at conducting a more recent and detailed analysis of the literature to help educators acquire a clearer understanding of the effectiveness of Philosophy for Children movement as *a globalized phenomenon*.

The Present Study

The purpose of the current meta-analysis is to examine the reported effectiveness of P4C from 2002 to 2016, immediately following the publication of the two articles that analyzed studies from 1970s to 2002. In addition, this meta-analysis examines which variables –participant age, socioeconomic status, study location, assessment measure, duration– of the intervention might moderate the magnitude of the aggregated effect sizes.

Through this meta-analysis, the researchers hope to find answers for the following question:

1. What does the cumulative research suggest regarding the overall effectiveness of P4C on students' cognitive abilities?
2. Do study design, students' backgrounds (grade level and socio-economic status), location and duration of intervention, and characteristics of cognitive outcome measurements influence the magnitude of the effect size of included studies?

Methodology

In this study, the effectiveness of P4C was tested through a meta-analysis, which is a method that merges the results of many independent researchers, conducted on a particular topic and performs statistical analysis (Çoğaltay & Karadağ, 2016).

Study Search and Retrieval

This study included the online databases British Education Index, ERIC, Education Full Text (H.W. Willson), Education Source, Academic Search Complete, PsycINFO database from 2002 to 2016. The keyword used was *philosophy n2 children*, which means it specified 2 maximum intervening words between philosophy and children, in any order. The researchers included both referred published journals and doctoral dissertations. Second, the researchers conducted a non-electronic journal search. The index of the journal *Thinking: Philosophy for Children* was consulted for articles. Then, potential relevant articles were retrieved from a library. The third was a google scholar search engine, *Journal of Philosophy in Schools*, as well as references listed in collected studies. Through the initial searches, 1180 articles were potentially relevant.

Inclusion Criteria

In order to be included in this meta-analysis, studies had to meet the following criteria:

1. Participants: The population of interest was pre-collegiate students enrolled in a Philosophy for Children program and their control-group counterparts. College studies and teacher education research were excluded from the study.
2. Intervention: Philosophy for Children has various names and diverse ways of practicing in the world. In this meta-analysis, we included studies that are under the names of P4C, Philosophy for Children, Philosophy with Children, and PwC. All the included studies must have explicit pedagogical markers of "community of (philosophical) inquiry" that shares the common practices of providing stimulus (stories, questions, pictures, or other media), students' questioning, and building on each other's ideas.

3. Publication: The retrieved study should be published between 2002 and 2016 in a *refereed journal* or as a thesis/dissertation.
4. Research design: 1) The study must be either random controlled experiments or quasi-experiments. 2) A quantitative measure of outcomes was used in the study to calculate the effect sizes of the intervention. 3) The outcome variables contained a measurement of cognitive outcomes, such as reasoning ability, comprehension ability, general cognitive ability, and academic development. 4) This meta-analysis focused on comparing the cognitive outcomes of P4C as the experimental group with other control groups where participants did not receive any thinking skill intervention. Thus, studies that did not contain a control group were excluded.

To ensure all studies included were well-designed and able to provide enough data for the computation of effect sizes, researchers left out studies that failed to conform to any of those criteria. Thus, a considerable number of studies were excluded in this stage particularly because many of them adopted qualitative methodology, which could not provide enough effect size for meta-analysis. This is because of the nature and limitation of meta-analysis itself as it applies only to research studies that produce quantitative findings. That is, studies using quantitative measurement of variables and reporting descriptive or inferential statistics to summarize the resulting data (Lipsey & Wilson, 2001). This process ruled out qualitative forms of research such as case studies, ethnography, and ‘naturalistic’ inquiry (Lipsey & Wilson, 2001).

Because of the large amount of literature, there were two steps of screening during the selection of included studies. First, one of the authors screened the titles and abstract of each of the 1,180 studies. 44 articles which met all the criteria of inclusion remained from the initial screening. Then, the full text of each of the 44 studies was retrieved and scrutinized. During this second screening of 44 full texts, 28 more studies failed to meet the requirements of the inclusion criteria. Finally, the remaining 16 studies were subject to the coding process.

Coding Procedure

The coding process is a data extraction process, picking clear and appropriate data from the pile of complex information (Çoğaltay & Karadağ, 2016). The manual for coding studies was developed by the researchers before proceeding to the coding.

Content validity for the coding sheet and coding manual was determined originally by submitting to scholars/researchers for feedback on the appropriateness of variables and categories created in this study. The first scholar works primarily on culture and curriculum, and the second professor is in educational psychology. There were two coders in this study. The first coder is the first author of this meta-analysis; the second coder is a doctoral student and the 3rd co-author. Both of the two coders have received statistical education for quantitative research.

After the coding manual was created, two coders initially met to go over the coding manual until everything was clear. The coders scrutinized each of the articles and extracted the variables and outcomes from the studies and input them into an excel document. To determine interrater reliability, the two researchers independently coded five studies (31.25% of the 16 articles) to ensure that the inclusion/exclusion criteria were met. The researchers achieved an interrater reliability of 90.0% across those studies. Analysis of coder disagreements resulted in the refinement of some definitions and decision rules for some codes. Then, each coder individually coded the remainder of the studies.

During the coding process, the first coder contacted the original authors from two different references for standard deviations and means to calculate the effect sizes. One set of data was obtained and another contact for data was not successful.

Data Analysis

Effect size computation, test of homogeneity and moderator analysis were conducted in the stage of data analysis. The effect size acquired in the meta-analysis study is a standard measure value used to determine the strength and direction (Çoğaltay & Karadağ, 2016) of the effectiveness of Philosophy for Children program on students' cognitive outcomes. Cohen's *d* was used to adjust and determine the effect sizes of each study. All the effect sizes in each study were aggregated to one effect size as the cognitive outcome. In meta-analysis, the unit of analysis is the individual research study, and any two or more effect sizes that come from the same study are statistically dependent (Lipsey & Wilson, 2001). Furthermore, all data analyses involving effect sizes were weighted analysis.

Two main models, namely fixed effects model and random effects model, were utilized in the analysis of heterogeneous distribution of effect sizes. Under a fixed effects model, an effect size observed in a study is assumed to estimate the corresponding population effect with random error that stems only from the chance factors associated with subject-level sampling errors in that study (Lipsey and Wilson (2001). If it is believed that the research is not equal in terms of functionality, and if generalizations through the estimated effect size are to be made for greater populations, then the model that should be used is the random effects model (Çoğaltay & Karadağ, 2016).

Last, the authors evaluated the data generated by the analysis and determined 1) whether P4C had a positive impact on students' cognitive abilities to address the first research question and 2) if characteristics of study design, students' background location and duration of intervention, characteristics of cognitive outcome measurements could be the moderator(s) of P4C on the effect size of cognitive abilities to answer the second research question.

Results

Six studies were excluded from the 16 research papers during the last stage of this analysis. One study was excluded because participants were younger than first grade (Säre et al., 2016) and this study only focuses on K-12 students. The study conducted by Othman and Hashim (2006) was excluded because the experiment's control group was still participating in another thinking skills intervention. This study compared P4C to another thinking program (the Reader Response Program). Thus, the control group is not neutral. The control groups in all the included studies were not under any thinking skills intervention. Two studies were not included due to the lack of means and standard deviations to calculate the effect sizes (Colom, Moriyon, Magro, & Morilla, 2014; Walker, Wartenberg, & Winner, 2013). Another one (Gorard et al., 2015) was not from a peer-reviewed journal and thus excluded from the study. One study was excluded because its outcome measure was spiritual development, which is not considered to be within the realm of cognitive outcomes (Abaspour, Nowrozi, & Latifi, 2015).

Studies Included

A total of 10 controlled experiments were included in this analysis, which together report the findings of eight independent studies and two follow-up studies. Among the ten studies, nine were articles from peer-reviewed journals, and one is a dissertation (Youssef, 2014). Table 1 provides an overview of the characteristics of each citation included in the synthesis. The sample sizes in these studies ranged from 28 to 540, representing 1,509 students from second grade in elementary to 9th graders in high schools. The sample sizes of studies were adjusted in following way:

First, if the sample size of one study in post-test was smaller than the pre-test due to the loss of participants, then the whole sample size of this study was coded according to the number of participants of the post-test. Second, in the case of one study with a follow-up study (Fair et al., 2015a; Fair et al., 2015b), the sample sizes of the independent studies were adjusted to the corresponding student groups with the follow-up studies. Thus, the overall sample sizes in this meta-analysis are smaller than those in the original literature.

Regarding outcomes measures, four studies (Fair et al., 2015a; Fair et al., 2015b; Topping & Trickey, 2007a; Topping & Trickey, 2007b) used the Cognitive Ability Test, a standardized test called CAT in America and CogAT in United Kingdom. It measures students' verbal, Quantitative and nonverbal reasoning abilities (Lohman et al., 2001). Two studies (Lam, 2012; Marashi, 2008) chose the The New Jersey Test of Reasoning Skills (NJTRS), which was specifically designed to measure reasoning skills. One study (Naderi, 2014) selected Abedi's Test of Creativity which was formulated to measure rate of creativity based on Torrance Tests of Creative Thinking (TTCT). Another study (Youssef, 2014) utilized a standardized test called The Test of Reading Comprehension (TORCH) to measure reading comprehension ability of the students. One study (Abbasi & Ajam, 2016) developed a questionnaire to test the educational progress of science lessons. To verify the validity of the measurement, the researchers included 20 in-service teachers of second grade to review the instrument. In the study by Tok & Mazı (2015), both the Reading Comprehension Test and Listening Comprehension Test were designed by the researchers. They were developed through the framework of predicted objectives for the reading and listening comprehension learning field in the elementary fifth grade Turkish Course Curriculum.

Overall Effectiveness of P4C

The overall effect size aggregated from the ten studies was 0.43 with a 95% confidence interval, ranging from 0.33 to 0.53. According to Cohen's Rule of Thumb (VanVoorhis & Morgan, 2007), the mean effect size represents that P4C has a moderate, positive overall cognitive effect for students who are in 2nd to 10th grade.

In this study, the homogeneity test was found to be statistically significant ($Q = 26.59$, $p < 0.01$), which means that there is more variability in effect sizes than would be expected from sampling error around the mean. Table 2 provides the overall results and omnibus test of this meta-analysis.

Results of Moderator Analysis

Since the homogeneity test was found to be statistically significant, a moderator analysis was used to find out the potential explanations for variance among effect sizes. In this meta-analysis, subgroup analysis was employed to detect moderating effects. Seven moderator variables were tested: grade level, socioeconomic status of students, location of studies, study design (random or quasi-experiments), total time of intervention, outcome measures, and type of outcomes. Table 3 provides a detailed statistical description of the result of moderator analysis.

In this study, two of the seven moderators revealed statistically significant effects. They were research location (two subgroups: Asia and western countries) and type of outcomes (three subgroups: general cognitive ability, reasoning skills and academic achievement). The tests of homogeneity indicated no statistical differences by grade levels, socio-economic status of participants, methods of group assignments, duration of the intervention, and outcome measures. The following is the detailed description of each subgroup analysis.

Grade Level. The included studies were divided into two categories in terms of the grade levels: 2 to 5 (k=4) and 6 to 10 (k=6). As seen in table 3, the aggregated effect size (Cohen's d) of studies which recruited grade 2 to 5 students was 0.51, and the average effect size of studies with grade 6 to 10 students was 0.42. Q_B was 0.75 ($p > .75$). From the results of this moderator analysis, no significant difference was found between effect sizes of studies according to the grade levels of the samples.

Table 1

Characteristics of Included Studies

Reference	Study Type	Location	Sample Size	Grade/Age Level	Outcome Measure	Effect Size	Variance
(Abbasi & Ajam, 2016)	Intervention	Iran	50	Second	Questionnaire of Educational Progress*	0.870	0.080
(Fair et al., 2015b)	Intervention	United States	177	Seventh	CogAT	0.590	0.020
(Fair et al., 2015a)	Follow-Up	United States	115	Seventh Grade-Two Years after	CogAT	0.570	0.030
(Lam, 2012)	Intervention	China	28	Secondary School First Grade	NJTRS	0.590	0.190
(Marashi, 2008)	Intervention	Iran	60	Eighth	NJTRS	1.100	0.070
(Naderi, 2014)	Intervention	Iran	60	High School First Grade	Abedi's Test of Creativity	1.190	0.070
(Tok & Mazı, 2015)	Intervention	Turkey	74	Fifth Grade	Reading Comprehension Test* and Listening Comprehension Test*	0.162	0.035
(Topping & Trickey, 2007a)	Intervention	United Kingdom	540	Ten-year-old students	CAT	0.25	0.01
(Topping & Trickey, 2007b)	Follow-Up	United Kingdom	183	Ten-Year-Old Students (Two Years After)	CAT	0.400	0.020
(Youssef, 2014)	Intervention	Australia	222	Sixth Grade	Reading Comprehension Test	0.340	0.020

Note: CogAT: Cognitive Ability Test (American Version); CAT: Cognitive Ability Test (United Kingdom Version); NJTRS: New Jersey Test of Reasoning Skills; *: Tests developed by researchers

Table 2

Overall Results and Omnibus Test of P4C Studies

	k	N	Median ES (d)	Fixed Effect		Random Effect		Q
				ES (d)	95% CI	ES (d)	95% CI	
P4C	10	1,509	.58	.43	[.33, .53]	.50	[.33, .66]	26.59**

Note: k = study size; N = total number of participants; CI = confidence interval; Q = omnibus test of homogeneity.

** p < .01

Socio-Economic Status of Students. In this sample of studies, we used two categories for the socio-economic status (SES) of the participants. The first group consisted of students who received free lunch, or were classified as 'economically disadvantaged' by the local districts. The second group of students are not identified as part of the free-lunch program, or was classified as from middle (or upper) class families. However, no significant heterogeneity in effect sizes was found between the two groups of students.

Research Location. This meta-analysis covers 5 studies conducted in the Western world and 5 in Asian countries. Five of the examined studies conducted in Asian countries: Iran, Turkey, and China. The other five studies come from Western countries: United Kingdom, Australia and the United States.

The first reason for doing this moderator analysis is because P4C as an educational movement has its roots in the Western philosophical traditions, which bring about the authors' uncertainty of the viability of its globalization. For example, P4C's exclusive emphasis on dialogue (Gregory, 2011), which is different from some non-Western philosophical practices such as contemplation, may negatively impact its implementation and effectiveness.

Another impetus for this analysis is that of the socio-political environments in these non-Western countries. For example, countries such as Iran and China might be more resistant to Western rhetoric and democratic schooling as a whole. In *Is Respecting Children's Rationality in Their Best Interest in an Authoritarian Context?*, the authors (Ghazinejad & Ruitenberg, 2014) argued that P4C implementation in Iran must balance the teaching of critical thinking and the protection of children's safety in their communities. Giving the consideration that P4C and its democratic educational ideals not only have conflicts with the extant educational systems but also may bring clashes between individual students and their communities, the authors wonder if the effectiveness of P4C will be influenced by these factors at all.

Through moderator analysis, a significant difference between the two groups was found (Q = 5.16, p < .05). The studies in Asian countries had higher effect sizes (d=0.69) than those studies conducted in Western countries (d=0.39). Yet, the five studies in Asian countries had significantly lower sample size (n=272) than the five studies in non-Asian countries (n=1237).

Study Design. To warrant the rigor of this meta-analysis, the authors set up stringent criteria for the inclusion of studies in which only random controlled trials and quasi-experiments were brought in the synthesis. From the moderator analysis, no significant difference was found between effect sizes of random controlled experiments and quasi-experiments which were included in this meta-analysis.

Duration. The authors of this study divided the literature into three subgroups based on the duration of interventions: 5 to 20 hours (k=4), 21 to 40 hours (k=3), and more than 40 hours (k=3). The result showed that none of the duration levels statistically varied from one another. Thus, there was no noteworthy difference between different levels of duration of intervention in the effectiveness of P4C on students' cognitive outcomes.

Outcome Measure: CAT or Non-CAT. Studies included were examined according to their outcome measures. Four studies using Cognitive Ability Tests were accepted as CAT subgroup; six studies applying other types of outcome measures were accepted as Non-CAT subgroup. No significant heterogeneity was found between these two subgroups.

Types of Outcomes. A significant difference among different types of outcomes was found (Q_B = 15.44, $p < .001$). The studies (Lam, 2012; Marashi, 2008) which tested the improvement of reasoning skills through P4C yielded the largest estimations (d=1.06), which is a large effect size. P4C used in improving general cognitive abilities (d=0.40), which is a moderate effect size. Reading comprehension ability (d=0.28) is a small effect size. This suggests that P4C has significant, positive influence on students' reasoning skills, and moderate effects on general cognitive ability and comprehension ability.

Summary

The first research question in this meta-analysis concerned the direction and magnitude of the effectiveness of P4C on students' cognitive ability. The studies analyzed here showed an overall positive medium effect size on cognitive outcomes in general.

The second question was whether and how the effectiveness of P4C differed significantly depending on the moderator variables. The moderator analysis found statistically significant results in regard to the study location and outcome types of these studies. No significant differences were found as to different grade levels, socio-economic statuses of participants, methods of group assignment, duration times of intervention, and cognitive measures. The results suggest that Philosophy for Children has a positive moderate influence on students' cognitive outcomes.

Discussion

Ten studies were included in this meta-analysis to determine the effects of Philosophy for Children program on students' cognitive abilities, and what characteristics of the intervention, students and outcomes measures could influence the magnitude of such effect.

The Overall Effectiveness of P4C

According to the findings of this meta-analysis, the Philosophy for Children program has shown a moderate, positive influence on students' cognitive outcomes. This result corroborates the previous literature on the program that states that P4C has a positive impact on students' various types of cognitive abilities (Fair et al., 2015a; Fair et al., 2015b; García-Moriyón et al., 2005; Topping & Trickey, 2007a, 2007b; Trickey & Topping, 2004).

The cognitive outcomes comprise general cognitive ability, reasoning skills, creative thinking abilities, educational progress in science, reading and listening comprehension abilities. Among all of these types of cognitive outcomes, the Philosophy for Children program has large aggregated positive effect on students' reasoning skills, while moderate influences on other cognitive domains. The previous P4C meta-analysis that focused on reasoning abilities (García-Moriyón et al., 2005) also indicated the positive impact of P4C on students' reasoning skills. However, as the number of studies selected in this meta-analytic review was limited, the interpretation of those aggregated results needs to be cautious.

Discussions about Findings between P4C and Students' Grade Levels

As stated in the results section, there was no statistically significant cognitive outcome in the effectiveness of P4C based on the grade levels of students. This result sheds lights on the question regarding P4C and students' age. Philosophy education is traditionally assumed to be appropriate for students no younger than secondary school age (Lipman & Sharp, 1978). But this moderator analysis indicates that both the studies with grade 2 to 5 students and the studies with grade 6 to 10 students benefited from P4C program (grade 2-5: d=0.51; grade 6-10: d=0.42). There was no statistically meaningful difference between the aggregated effect sizes of the two subgroups.

Table 3

Moderator Testing of Study

Variable	k	*N*	d	95% CI	Q_B	ANOVA
Research Location						
Asia	5	272	.69	[.46, .91]		>N-A
Non-Asian Countries	5	1,237	.39	[.27, .51]	5.16*	
Grade at Intervention						
2-5	4	416	.51	[.34, .69]		
6-10	6	1,093	.42	[.29, .55]	0.75	
SES of Participants						
Disadvantaged	4	1,015	.40	[.27, .53]		
Others	6	494	.55	[.37, .72]	1.74	
Methods of Group Assignment						
Random	4	811	.44	[.33, .54]		
Quasi Experiment	6	698	.52	[.38, .66]	0.83	
Total Time of Intervention						
5-20 Hours	4	445	.34	[.18, .51]		
21-30 Hours	3	579	.28	[.13, .43]		
More than 40 Hours	3	427	.47	[.28, .66]	2.41	
Outcome Measure						
CAT or CogAT	4	1,015	.40	[.27, .53]		
Others	6	494	.55	[.37, .72]	1.74	
Type of Outcomes						
General Cognitive Outcomes	4	1,015	.40	[.27, .53]		
Reasoning Skills	2	148	1.06	[.72, 1.40]		> C & R
Reading Comprehension	2	296	.28	[.06, .50]	15.44***	

Note: k = study size; N = number of participants; CI = confidence interval; Q_B = between-groups test of homogeneity;

ANOVA = significant result. * $p < .05$, *** $p < .0$

Discussions about Findings between P4C and Locations

Another significant finding of this study is where the research was conducted could determine its effect sizes. A statistically significant difference was found between the effect sizes of studies in Western (d = 0.39) and non-Western (d=0.59) countries. From this result, it seems that the globalizing Philosophy for Children program has generated more positive influences on students' cognitive outcomes in non-Western countries than Western countries. This is not expected by the authors in that considering P4C's Western philosophical heritage, this program might not be suitable to the socio-political and philosophical contexts of non-Western educational settings and thus has less positive outcomes. There are several possible accounts for this phenomenon.

First, the studies in Asia have smaller sample sizes. Because P4C is still new to educators and researchers in those countries (Lam, 2012; Marashi, 2008), including Iran, China and Turkey, these studies are often pilot studies with small sample sizes. Moreover, since P4C was initiated in the United States in the 1970s (Brandt, 1988), it is more relatively well-known to the educators in the United States, United Kingdom, Australia and other Western countries. Thus, studies conducted in these areas tended to evaluate P4C in *large school districts* (Fair et al., 2015b; Toppings & Trickey, 2007a; Youssef, 2014). In this meta-analysis, the mean sample size of Western studies is three times higher than the mean sample size of non-Western studies. Smaller sample sizes may contribute to the quality of teacher education and P4C implementation. Pedagogically speaking, all the studies have utilized community of inquiry as the core pedagogy. Though it might be the case that there are more experienced practitioners in countries where P4C is more well-known or in studies with larger scale, it is also possible that the reverse is true. This is because the practitioners in pilot studies may have received more focused teacher education while in studies with large sample size teacher education and motivation for practicing P4C are not easy to control.

Another possible explanation is that several studies in non-Western countries tested the improvement of reasoning skills among students (Lam, 2012; Marashi, 2008; Othman & Hashim, 2006), while no Western research included here specifically examined the reasoning abilities of students. According to the moderator analysis regarding the effect sizes of studies with different types of outcomes, there is a statistically significant difference between reasoning skills and other types of outcomes. If P4C is more effective to the improvement of reasoning skills, then the discrepancy between the effect sizes in Western and non-Western studies is understandable.

Although none of the studies from non-western countries in this meta-analysis seemed to utilize or create P4C philosophical texts that catered to their specific philosophical, socio-political, and educational environments, it is worthwhile to consider the complex and nuanced cultural and political consequences of introducing P4C to non-Western countries from the lens of de-colonial theories and democratic education (Ghazinejad & Ruitenberg, 2014; Ndofirepi & Cross, 2015), and also the possibility of P4C being transformed by its globalization (Gregory, 2011).

Discussions about Findings between P4C and Duration of Interventions

The moderator analysis showed that the P4C's influence was not moderated by the duration of the intervention. This was not expected since several studies (Fair et al., 2015a; Fair et al., 2015b; García-Moriyón et al., 2005; Topping & Trickey, 2007a) have proposed that P4C should be implemented for a significant period of time before the program shows results. That being said, our result resonates with one P4C study (Fair et al., 2015b). In this project, the authors replicated a previous experiment conducted by Topping and Trickey (Topping & Trickey, 2007a), in which they shortened the duration of the P4C intervention to less than half of the former one: from 58 weeks to 22 to 26 weeks. The result showed that P4C still had a moderate effect on students' general cognitive ability. This suggests that a short time of exposure to P4C may also have a meaningful impact on students' cognitive outcomes.

Discussions about the Excluded Literature

A large number of studies were excluded in the process of the analysis. The first reason is because the majority of studies in the field of P4C are qualitative and theoretical, whereas the methodology employed in this study is a quantitative meta-analysis that needs to extract the data from many independent studies conducted on a particular topic and perform statistical analysis (Çoğaltay & Karadağ, 2016). For example, numerous insightful articles regarding P4C have been published in Africa (Di Masi & Santi, 2016; Ndofirepi & Cross, 2015), but none of them was quantitative and could be used in this study.

The second reason is that even if some studies utilized a quantitative methodology, they often lacked sufficient information especially for the means and standard deviations for the researcher to compute effect sizes.

Third, the result of exhaustive literature search and process of study inclusion/exclusion showed that more rigorous quantitative studies regarding P4C program are still needed. The researchers gathered over 1180 studies at first, after coding procedure, there were only 16 studies remained. Throughout the data analysis process, six more articles were excluded from the study. The main reason of this phenomenon is that the majority of the literature regarding P4C is qualitative and theoretical. Due to the nature of meta-analysis, which is a quantitative synthesis study, it cannot process and analyze qualitative and non-empirical literature (Lipsey & Wilson, 2001). In addition, there is not only few quantitative experiments in P4C, but the data produced by the studies are not sufficient enough for computing an effect size. Thus, this suggests that this field needs more studies to form a larger cluster of rigorous research.

Moreover, some studies gave a novel practice and detailed observation of children who are below the age of five. To narrow down the age to grade 1-12 students, one study about P4C's effectiveness on reasoning skills was excluded, but it definitely shows the potential of teaching and introducing philosophy to very young children. For example, the study conducted by Dasi et al. (2013) showed a clear significant improvement in socio-psychological abilities among the 5-year-old children and a partial improvement in the 4-year-old children after participating in a few sessions of the P4C program. Also, one study (Säre et al., 2016) showed that P4C positively influenced preschoolers' verbal reasoning skills. These studies provide information for educators and researchers to understand the unfamiliar area in which young children are involved in rather than exclude them from this philosophy. Future research can consider examining how P4C affects the cognitive outcome of children in kindergarten or preschool. Types of P4C study outcomes can be expanded from cognitive outcomes to psychological or social outcomes.

Conclusion and Suggestions for Future Research

P4C was found to have a moderate, positive overall effect on students' cognitive outcomes. The authors suggest that P4C may be considered as an effective thinking program for teachers in grade 2-10 education. Based on the findings of this meta-analysis, several recommendations and suggestions for future research are advanced:

First, in addition to long-term implementation of P4C in classroom, a short time of exposure to P4C may also have a meaningful impact on students' cognitive outcomes. The practice of P4C should not only be limited to the realm of long-term applications.

Second, this study suggests that grade level is not a moderator of the effectiveness of P4C in improving students' cognitive abilities. Moreover, a small number of studies (Dasí et al., 2013; Säre et al., 2016) have practiced P4C with very young children who are below the age of five. Thus, age should not be the sole reason for excluding students from philosophy education, and more studies are needed in terms of the impacts of P4C on very young children. As Lone and Burroughs have said (2016), at one time or another we all ask philosophical questions of some kind, consider our values and reflect on the rightness and wrongness of our actions. It is possible that all children, regardless of age and grade level, have the capacity and interest to engage in philosophical activities (Lipman, 2009).

Lipman and Sharp (1978) once questioned a presupposition of P4C which assumes philosophy education should be assigned to students who are either from gifted programs or from particular advantageous backgrounds. The results of this study also indicate that there is no statistically significant difference in the impact of P4C based on social-economic backgrounds. It is suggested that educators in P4C program should strive to build a community of inquiry that encourages students to share not only divergent social backgrounds and life experiences (Lipman, 2009) but also different styles of thinking (Lipman & Sharp, 1978) so as to involve them in the classroom discussion.

Different from previous meta-analyses, this study emphasized the exhaustive search for P4C studies around the world. The results show that during recent years, a considerable amount of practices have been taken in various continents. This study calls for more research and analyses that consider the nuances and details of P4C practices in different cultural, social, educational, linguistic, and philosophical contexts.

Last, while P4C as a famous thinking skill program has been relatively examined (Daniel & Auriac, 2011; Fair et al., 2015a; Fair et al., 2015b; García-Moriyón et al., 2005; Säre et al., 2016; Topping & Trickey, 2007a, 2007b), limited research on socio-psychological outcomes have been generated in this field. More studies that explore the connection between community of inquiry, philosophical thinking and the socio-psychological development of children are strongly recommended.

References

Abaspour, N., Nowrozi, R. A., & Latifi, Z. (2015). Investigating the Effect of Educating Philosophy in the Children on the Spiritual Development of Female Students with 12-14 Years Old in the City of Isfahan. *Journal of Education and Practice, 6*(11), 162-166.

Abbasi, Z., & Ajam, A. A. (2016). The Effects of Philosophical Stories on Emotional Intelligence and Educational Progress of Students in Science Lessons. *Mediterranean Journal of Social Sciences, 7*(2), 282.

Brandt, R. (1988). On Philosophy in the Curriculum: A Conversation with Matthew Lipman. *Educational Leadership, 46*(1), 34-37.

Butnor, A. (2012). Critical Communities: Intellectual Safety and the Power of Disagreement. *Educational Perspectives, 44*, 29-31.

Çoğaltay, N., & Karadağ, E. (2016). The Effect of Educational Leadership on Organizational Variables: A Meta–Analysis Study in the Sample of Turkey. *Educational Sciences: Theory & Practice, 16*(2).

Colom, R., Moriyon, F. G., Magro, C., & Morilla, E. (2014). The Long-term Impact of Philosophy for Children: A Longitudinal Study (Preliminary Results). *Analytic Teaching and Philosophical Praxis, 35*(1).

Daniel, M., & Auriac, E. (2011). Philosophy, Critical Thinking and Philosophy for Children. *Educational Philosophy and Theory, 43*(5), 415-435.

Dasí, M. G., Quintanilla, L., & Daniel, M. F. (2013). Improving Emotion Comprehension and Social Skills in Early Childhood Through Philosophy for Children. *Childhood & Philosophy, 9*(17), 63-89.

Dewey, J. (1916). *Democracy and Education: An Introduction to Philisophy of Education*: Macmillan.

Dewey, J. (1986). *Experience and Education.* Paper Presented at the The Educational Forum.

Di Masi, D., & Santi, M. (2016). Learning Democratic Thinking: A Curriculum to Philosophy for Children as Citizens. *Journal of Curriculum Studies, 48*(1), 136-150.

Fair, F., Haas, L. E., Gardosik, C., Johnson, D., Price, D., & Leipnik, O. (2015a). Socrates in the Schools: Gains at Three-year Follow-up. *Journal of Philosophy in Schools, 2*(2).

Fair, F., Haas, L. E., Gardosik, C., Johnson, D. D., Price, D. P., & Leipnik, O. (2015b). Socrates in the Schools From Scotland to Texas: Replicating a Study on the Effects of a Philosophy for Children Program. *Journal of Philosophy in Schools, 2*(1).

García-Moriyón, F., Rebollo, I., & Colom, R. (2005). Evaluating Philosophy for Children. *Thinking: The journal of philosophy for children, 17*(4), 14-22.

Ghazinejad, P., & Ruitenberg, C. (2014). Is respecting children's rationality in their best interest in an authoritarian context? Ethics and Education, 9(3), 317-328.

Giménez-Dasí, M., Quintanilla, L., & Daniel, M.-F. (2013). Improving emotion comprehension and social skills in early childhood through philosophy for children. childhood & philosophy, 9(17).

Gopnik, A. (2009). *The Philosophical Baby: What Children's Minds Tell us About Truth, Love & the Meaning of Life*: Random House.

Gorard, S., Siddiqui, N., & Huat See, B. (2015). Philosophy for Children: Evaluation Report and Executive Summary. *Education Endowment Foundation, Millbank, UK.*

Gregory, M. (2011). Philosophy for Children and Its Critics: A Mendham Dialogue. *Journal of Philosophy of Education, 45*(2), 199-219.

Gregory, M. R., & Laverty, M. J. (2017). In Community of Inquiry with Ann Margaret Sharp: Childhood, Philosophy and Education: Routledge.

Hidi, S., Renninger, K. A., & Krapp, A. (2004). Interest, A Motivational Variable That Combines Affective and Cognitive Functioning. *Motivation, Emotion, and Cognition: Integrative Perspectives on Intellectual Functioning and Development*, 89-115.

Jackson, T. (2001). The Art and Craft of 'Gently Socratic' Inquiry. *Developing Minds: A Resource Book for Teaching Thinking, 3*.

Lam, C.-M. (2012). Continuing Lipman's and Sharp's Pioneering Work on Philosophy for Children: Using Harry to Foster Critical Thinking in Hong Kong Students. *Educational Research and Evaluation, 18*(2), 187-203.

Lipman, M. (2004). Philosophy for Children's Debt to Dewey. *Critical and Creative Thinking, 12*(1), 1-8.

Lipman, M. (2009). (USA) Philosophy for Children: Some Assumptions and Implications. *Children Philosophize Worldwide: Theoretical and Practical Concepts, 9*, 23.

Lipman, M., & Sharp, A. M. (1978). Some Educational Presuppositions of Philosophy for Children. *Oxford Review of Education, 4*(1), 85-90.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-analysis* (Vol. 49): Sage Publications Thousand Oaks, CA.

Lohman, D. F., Thorndike, R. L., Hagen, E. P., Smith, P., Fernandes, C., & Strand, S. (2001).
Cognitive Abilities Test third edition. London: nferNelson.

Lone, J. M., & Burroughs, M. D. (2016). *Philosophy in Education: Questioning and Dialogue in Schools*: Rowman & Littlefield.

Lone, J. M., & Green, M. (2013). Philosophy in High Schools: Guest Editors' Introduction to a Special Issue of Teaching Philosophy. Teaching Philosophy, 36(3), 213-215.

Love, R. (2016). The Case for Philosophy For Children in the English Primary Curriculum. *Analytic Teaching and Philosophical Praxis, 36*(1).

Marashi, S. M. (2008). Teaching Philosophy to Children: A New Experience in Iran. *Analytic Teaching, 27*(1), 12-15.

Matthews, G. B. (1980). *Philosophy and the Young Child*: Harvard University Press.

Matthews, G. B. (1994). *The Philosophy of Childhood*: Harvard University Press.

Meskin, A., & Cook, R. T. (2012). The Art of Comics: A Philosophical Approach: John Wiley & Sons.

Murris, K. (1992). Teaching philosophy with picture books: Infonet.

Ndofirepi, A., & Cross, M. (2015). Child's Voice, Child's Right: Is Philosophy for Children in Africa the Answer? *Interchange, 46*(3), 225-238.

Ndofirepi, A. P., & Shanyanana, R. N. (2016). Rethinking ukama in the context of 'Philosophy for Children'in Africa. Research Papers in Education, 31(4), 428-441.

Nia, A. T. (2014). Investigate the Effect the Philosophy for Children Program (p4c) on Reducing Trait Anger in Teens. *Stud, 4*(2), 449-455.

Othman, M., & Hashim, R. (2006). Critical Thinking & Reading Skills. *Thinking: The Journal of Philosophy for Children, 18*(2), 26-34.

Peirce, C. S. (1868). Some Consequences of Four Incapacities. *The Journal of Speculative Philosophy, 2*(3), 140-157.

Peirce, C. S., & Houser, N. (1998). *The Essential Peirce: Selected Philosophical Writings* (Vol. 2): Indiana University Press.

Piaget, J. (1931). Children's Philosophies.

Roberts, A. F. (2006). The effects of a teacher development programme based on Philosophy for Children. University of the Western Cape.

Säre, E., Luik, P., & Tulviste, T. (2016). Improving Preschoolers' Reasoning Skills Using the Philosophy for Children Programme. *Trames: A Journal of the Humanities and Social Sciences, 20*(3), 273.

Scholes, L., Lunn Brownlee, J., Walker, S., Johansson, E., Lawson, V., & Mascadri, J. (2016). Promoting Social Inclusion in the Early Years of Elementary School: A Focus on Children's Epistemic Beliefs for Moral Reasoning. *International Journal of Inclusive Education*, 1-14.

Sharp, A. M., Reed, R. F., & Lipman, M. (2010). *Studies in Philosophy for Children: Harry Stottlemeier's Discovery*: Temple University Press.

Tok, Ş., & Mazı, A. (2015). The Effect of Stories for Thinking on Reading and Listening Comprehension: A Case Study in Turkey. *Research in Education, 93*(1), 1-18.

Topping, K. J., & Trickey, S. (2007a). Collaborative Philosophical Enquiry for School Children: Cognitive Effects at 10–12 Years. *British Journal of Educational Psychology, 77*(2), 271-288.

Topping, K. J., & Trickey, S. (2007b). Collaborative Philosophical Inquiry for Schoolchildren: Cognitive Gains at 2-year Follow-up. *British Journal of Educational Psychology, 77*(4), 787-796.

Trickey, S., & Topping, K. J. (2004). 'Philosophy for Children': A Systematic Review. *Research Papers in Education, 19*(3), 365-380.

van der Straten Waillet, N., Roskam, I., & Possoz, C. (2015). On the Epistemological Features Promoted by 'Philosophy for Children'and Their Psychological Advantages When Incorporated into RE. *British Journal of Religious Education, 37*(3), 273-292.

Vansieleghem, N., & Kennedy, D. (2012). Philosophy for children in transition: problems and prospects (Vol. 15): John Wiley & Sons.

VanVoorhis, C. W., & Morgan, B. L. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology, 3*(2), 43-50.

Walker, C. M., Wartenberg, T. E., & Winner, E. (2013). Engagement in Philosophical Dialogue Facilitates Children's Reasoning about Subjectivity. *Developmental Psychology, 49*(7), 1338.

Youssef, C. (2014). A Multilevel Investigation into the Effects of the Philosophical Community of Inquiry on 6th Grade Students' Reading Comprehension, Interest in Maths, Self-esteem, Pro-social Behaviours and Emotional Well-being (Doctoral Dissertation). *Queensland University of Technology, Australia.*

Address Correspondences to:

Sijin Yan (Correspondence Author), Texas A&M University, Teaching, Learning & Culture Department, Philosophy Department. 321 YMCA Building, College Station, TX 77843-4237. Email: yansj101@tamu.edu. Tel: 979-676-7911

Lynne Masel Walters, Texas A&M University, Teaching, Learning & Culture Department, 362 Harrington Office Building, College Station, TX 77843 Email: lynne-walters@tamu.edu Tel: 979-845-8384

Zhuoying Wang, Texas A&M University, Texas A&M University, Teaching, Learning & Culture Department, 111D Harrington Office Building, College Station, TX, 77843 Email: ustop2013wzy@tamu.edu Tel: 979-587-4603

Chia-Chiang Wang, Department of Counseling, School, and Educational Psychology, University at Buffalo, 413 Baldy Hall, Buffalo, NY, 14260-1020 Email: chiachia@buffalo.edu Tel: 716-645-1119