An Evaluation of the 'Philosophy for Children' programme: The impact on Cognitive and Non-Cognitive Skills

OURANIA MARIA VENTISTA

Durham University
School of Education

An Evaluation of the 'Philosophy for Children' programme: The impact on Cognitive and Non-Cognitive Skills

Ourania Maria Ventista

PhD Thesis

School of Education
Durham University

2019

Supervised by: Prof Stephen Gorard and Dr Nadia Siddiqui

To my angel,
Kostas Eleftheriadis

Contents

Statement of Copyright

Acknowledgements

I would like to thank my supervisors Prof Stephen Gorard and Dr Nadia Siddiqui who supported me during this journey. During supervision, they asked the right questions and taught me how to lead my research project.

I would like to thank the head teachers, teachers and students who participated in this research. This research was conducted thanks to their help.

I would also like to thank Prof Stuart Kime. By challenging me, he helped me discover some of my strengths (zone of proximal development!).

I would also like to thank Hockerill Educational Foundation for annual grants which contributed to my PhD maintenance.

I would like to thank Prof Giasemi-Olga Sarafidou and Dr Marita Paparoussi. They played the biggest role in my decision to pursue a PhD. I would like to say thank you for introducing me to educational research and for your invaluable support.

I would like to thank Grigoris Arkoumanis. Even though he met me at the end of this journey, he supported my dream.

I would also like to thank my mum (Sofia Spartali) for always being next to me and my dad (Giorgos Ventistas) for his help with my studying and for teaching me to be independent. I would like to thank my grandparents (Ourania and Konstantinos Ventistas, Chrisanthi and Konstantinos Spartalis) for educating me all these years. I would like to thank Efi Lazaridou and Giannis Vassiliadis for believing in me. Last but not least, I would like to thank my three siblings; Aggelos Ventistas, Konstantina Ventista and Konstantina Vassileiadi (who appear on my PhD cover). Their natural curiosity and their presence in my life have been an inspiration for me and my research.

Abstract

Philosophy for Children (P4C) is a school-based intervention currently implemented in more than 60 countries. This thesis examines the evidence regarding the effectiveness of Philosophy for Children for developing pupils' cognitive and non-cognitive skills.

Three different approaches were used. A systematic literature review was conducted of the evidence published in the last 40 years. A new comparative evaluation study was conducted with Year 5 pupils in 17 primary schools in England (N = 547 pupils in the intervention group, N= 270 in the comparison group). The intervention lasted for an academic year, and a pre-test and a post-test were given at the beginning and end of the school year to evaluate students' critical thinking and creativity. Secondary data analysis of the National Pupil Database (NPD) from the Department of Education was used to examine the long-term effect of P4C implementation on attainment (reading, writing, maths). The results of 34 schools which implemented P4C during Key Stage 2 (2011-2015) were compared with 14,791 mainstream schools in England which did not, and the same analysis was repeated based only on these pupils in both groups known to be eligible for Free School Meals during the last six years (as an assessment of the impact of the P4C on narrowing the poverty attainment gap).

The review results suggested that P4C generally has a positive impact on reasoning skills. In most studies, P4C also has a positive impact on literacy and some non-cognitive skills. However, the new comparative evaluation study found no evidence that P4C has a positive impact on Year 5 students' critical thinking or creativity. This comparative study has some limitations in terms of design and inevitable attrition. The more robust secondary data analysis showed that students eligible for Free School Meals develop their reading and writing more after long-term P4C implementation than in non-P4C schools, during Key Stage 2.

By combining all of the evidence from the review, comparative evaluation study and secondary data analysis, this study suggests that the implementation of P4C in primary schools is still worthwhile, both in its own terms and for its added benefits in terms of cognitive and perhaps non-cognitive outcomes. The programme is likely to help improve students' reasoning skills. P4C can improve the literacy of disadvantaged students in the classrooms, relative to their peers, and so contribute towards closing the

attainment gap between advantaged and disadvantaged students. However, the new evaluation creates a caution about what can be expected from P4C and, if it used, the programme may need adjusting in order to provide opportunities for practicing a wider range of thinking skills.

1. Introduction

In this introduction, I explain the reasons why the research study of this thesis can be considered of interest for educators and educational policies which aspire to the cultivation of the cognitive and non-cognitive skills of primary school students. I also present the four research questions for this thesis, and conclude by presenting the thesis outline.

1.1. Purpose of the study

This thesis examines the impact of the Philosophy for Children (P4C) programme mainly in primary education. P4C does not teach specific content and can be considered a skills-based intervention. Although the importance of the development of thinking skills is prioritised, the evaluation of the programme in this thesis is not focused only on this or on a specific subject. A holistic and multi-dimensional evaluation of the programme is discussed. This thesis rather examines the contribution that P4C can play in the development of cognitive and non-cognitive skills. There are many economic and social reasons which make the teaching of thinking in schools of great importance. As society experiences changes schooling is required to adapt to the needs of the changing economy and society (Jones & Idol, 1990).

Traditional education is more focused on the transmission of existing knowledge which is organised into subjects, whilst progressive education more on the needs and interests of the learner for what is going to be learnt (Pring, 2007). This debate underlies how education perceives truth and knowledge and therefore what should be taught in schools. Traditional education appears to promote the status quo while progressive education is perhaps more future-orientated. Mitra (2000) said that current knowledge will no longer be valid when the pupils leave school, and the students will have to create new paradigms for new problems in their later life.

The term '21st-century learning' refers to education which prepares students for the socioeconomic and political context characterised by globalisation and ever-changing digital technology (Benade, 2014). Even though it could be questioned to what extent education should prioritise the employer's requirements, it is probably acceptable that schooling should be sensitive to these needs. For example, there has been a decrease in the need for activities involving manual labour and an increase in the need for cognitive activities.

Hence, I argue that the debate should focus on what type of knowledge should be taught, or what '21st-century learning' should involve? Should schools teach students existing knowledge or develop skills? Hirsch (2011) suggested that core knowledge is required either way and argued in favour of the knowledge taken for granted in classroom and society. Therefore, he suggested a curriculum that builds knowledge grade by grade in specific disciplines, such as maths and science. I suggest that recall, rote learning and memorization should stop being the centre of education. I suggest that schools should not implement knowledge-based curricula with the traditional sense.

I associate knowledge-based curricula with hidden curricula. A hidden curriculum might promote particular knowledge, work-related behaviour, such as conforming to authority (LeCompte, 1978) and reproduce the status quo. Learners should also be equipped with the skills to evaluate this knowledge. I argue that in progressive education the freedom that the learner receives is less likely to promote aims of hidden curricula because teaching thinking can either be neutral content or even transformative by promoting independent thinking.

Nowadays, access to information is relatively easy and the amount of information available online is enormous and growing. Individuals still need knowledge, but they mainly need the skills to be able to search for information, judge its trustworthiness, and process it in an appropriate manner. In a sense, it is the trustworthiness of data as judged by the user that makes it 'information' rather than just noise. Due to fast changes in society and economy, it has been suggested that the knowledge demands for the 21st century are not easily predictable, and thus the education system of each country needs to foster critical thinking and creativity (Berliner, 2011).

Furthermore, the era of the 21st century has been characterised as a post-truth era. Reznitskaya and Wilkinson (2017) recognised that many people appear to think that there is no objective knowledge. For instance, in history there is the idea of historic scepticism and relativism, where historians cannot agree about what happened and there is no objective history (Blake, 1955). If there is no objective knowledge, then there is no need for knowledge-based curricula. Hence, if relativism is accepted, the demand for teaching thinking skills, such as creativity, critical thinking and problem solving, is crucial.

This should not mean that knowledge should be completely disregarded. Pring (1980) argued that schooling should develop the mind and he suggested that knowledge is necessary for the development of mind. According to Pring, knowledge should focus both on 'knowing that' (propositional knowledge) and 'knowing how'. This thesis is mainly focused on 'knowing how' in education. This type of knowledge in this thesis can still be related to the main subjects, such as literacy and maths. Students should be taught how to write, how to read, how to do maths, how to think.

It might be questionable to what extent the schooling succeeds in developing the thinking skills of the students in practice. A few decades ago Lipman identified a problem in schooling. Young children start school with a natural curiosity. However, the schooling does not effectively develop this curiosity. Lipman (1976, p.22) argued that the school aggravates the thinking of the students instead of expanding it;

> What the school does succeed in introducing into the child is a negative charisma, a gratuitous belief in his own intellectual impotence, a distrust of any intellectual powers of his own other than what it takes to cope with problems formulated and assigned to him by others. The lively curiosity that seems to be an essential part of the child's natural impulse is sooner or later beater or battered out of him by the intransigencies of the educational system.

In other words, it can be disputed whether education in fact supports the pupils' thinking development. As a solution to this apparently ineffective aspect of schooling, Lipman designed the Philosophy for Children programme and he explicitly claimed that the programme improved the critical thinking and creativity of students (Lipman, 2003).

P4C adherents claimed that implementing P4C in primary schools increases students' creativity and critical thinking (Fisher, 2003; Lipman, 1976; 1995; 2003). It is still unclear whether the existing evidence supports this claim. The impact on some areas still remains unexamined and the evidence for others is contradictory, as it will become evident later in this thesis. Despite the lack of a coherent discussion of all the available evidence, the programme is very popular and it is currently implemented in approximately 60 countries (SAPERE, 2015a). However, according to an evidence-based education, school policy and practice should be justified based on sound evidence (Coe, 1999). Education in England should be evidence-informed which means that the policy-makers and practitioners should base their decisions on evidence about

effectiveness. It is useful to know whether a programme works before time and money are spent on its implementation and therefore it is important to examine and combine the available evidence regarding the programme effectiveness.

In recent years the Education Endowment Foundation (EEF) funded trials to produce evidence with main interest to reduce the attainment gap between poor students and their peers (Education Endowment Foundation, 2018a). However, it also contributed to an increase of interest in educational evidence about effective school interventions since the launch of the Teaching and Learning Toolkit in 2011. This toolkit became very popular among teachers in England and shed light on effective interventions, their impact and costs, according to its creators (Higgins et al., 2016). The Toolkit 'helped to create a more evidence-led culture in the classroom' (Higgins, 2017). The creation of Research Schools in England, the popularity of events such as ResearchED among teachers and their participation in discussions in social media, such as Twitter, show their willingness to implement evidence-based practice. Another example is the creation of the Durham University Evidence Centre for Education which aims to provide and synthesise evidence to inform policy and school practice.

Consequently, the purpose of this thesis is to explore, generate, synthesise and evaluate the evidence regarding the impact of P4C on cognitive and non-cognitive skills. This study offers an overall evaluation of the programme and it investigates whether this programme is effective and worth implementing in the school classrooms in England. Even though this thesis presents existing bibliographic evidence coming from around the world, the primary and secondary data analyses are based on English schools. Therefore, the study mainly discusses P4C effectiveness in English primary schools. The results, however, could be a useful indicator for evidence-based policies in other countries.

1.2. Research Questions

Having presented the purpose of the study, it is important to present the research questions of the study. The four research questions set are:

a) According to the current published evidence, is P4C effective in improving students' cognitive and non-cognitive skills?

b) Does P4C programme have an impact on the critical thinking of Year 5 students in primary schools in England?

c) Does P4C programme have an impact on the creativity of Year 5 students in primary schools in England?
d) Does P4C have an impact on students' attainment during Key Stage 2?

These questions are addressed through three different approaches. The first one via the conduct of a systematic literature review of the evidence published in last thirty-five years. The second and third questions are examined by a quasi-experimental trial in primary schools in England. Finally, the fourth one is examined based on an analysis of secondary data from the Department for Education (DfE) in England. All the questions of this thesis can be summarised in one big question, which refers to the programme effectiveness in primary schools in England overall. This thesis investigates multi-facet indicators of programme effectiveness to conclude in favour of or against its wider implementation in schools.

1.3. The Significance of the study

This study is significant because of its focus on a popular school-based intervention, and crucial skills that education aims to develop. First, the study examined the impact of P4C, which is currently implemented in approximately 60 different countries (SAPERE, 2015a), and provides a multidimensional evaluation of the programme. According to an evidence-based approach to education, the programmes implemented in schools should be trialed for their effectiveness (Coe, 1999). This study examined whether the time, effort and money spent on the programme has some impact on the various skills of the pupils. Previously published studies examined the impact of the programme on a specific skill. There was no recent study which combined all the available evidence to lead to evidence-informed practice. Hence, this study is significant because it contributes towards informed decisions about the implementation of the programme in schools.

Secondly, the study examined P4C impact on critical thinking and creativity. Teaching thinking is important because it might have an impact on pupils' life at the time and later in life. Critical thinking skill can be considered important if it helps people's actions to be informed and compatible to desirable outcomes and they can avoid being a victim of propaganda or hegemony (Brookfield, 2012). In this era, thinking skills and questioning can support the students in the quest for truth and

meaning. Therefore, as it has already been explained, thinking skills will be useful for pupils in the cases of relativism or propaganda.

What is more, this study is meaningful because it examined whether P4C can develop creativity in primary schools in England. The development of creativity is an issue which should be taken seriously, since creativity may contribute to the health of the person, both physically and psychologically, adaptability, self-expression, and problem solving (Runco, 2004).

The development of thinking skills is currently targeted by educational curricula across the world. For example, the national curriculum in Australia urges the development of seven general capabilities, which include creative and critical thinking (Australian Curriculum Assessment and Reporting Authority, n.d.). It is not surprising that Wyse and Ferrari (2015), who examined European national curricula of twenty-seven countries, identified that creativity is an educational aim included in all of them. The significance and emphasis attributed might vary within each curriculum, but creativity is an educational target in all.

Whilst creativity is an important educational target in curricula across the world, its development is usually restricted to the Arts subjects. The English curriculum for primary education (Wyse & Ferrari, 2015) currently underestimates its significance. Even though the development of creativity can take place across all subjects of the curriculum, in the guidelines of the national curriculum creativity was mostly emphasised in the arts subjects. They found a difference between the occurrence of the word creativity in the primary and secondary curriculum in England. The primary school curriculum texts referred much less to the development of creativity. This finding does not necessarily suggest that creativity is developed less in primary than secondary schools in England. However, it suggests that the national curriculum for primary education does not prioritise the development of this skill to the same extent as the secondary education curriculum. As a result, since creativity is an educational aim and there is a gap in the curriculum in primary schools in England, this study could potentially provide some evidence for the development of this skill in this educational stage.

I consider this focus on thinking skills beneficial for various reasons. I have already discussed how this focus can be appropriate for the current economic and social situation in the 21st century. Furthermore, I argue another potential benefit of making a shift towards thinking skills can be linked to the attainment gap between

advantaged and disadvantaged students. The attainment gap is an existing reality in education. However, other types of assessments can be less susceptible to influences from socioeconomic status. For example, the assessment of creativity can be fair way without discriminating particular groups of students. Specifically, the assessment of creativity has been found to compensate ethnic, race and gender differences, which are common in the performance of the students in IQ and attainment tests (Kaufman & Sternberg, 2007). Consequently, without making a strong claim about it, assessing thinking skills and prioritising them in education might provide a fairer education. Thus, this study is significant because it emphasises these skills.

Finally, this thesis is important because it also focuses on the impact of the programme on attainment. Attainment has been suggested as an important predictor of the wellbeing of the students in later life, since attainment is linked to the income, the health and the socio-emotional wellbeing they have as adults (Child Trends, 2016). Schools are currently accountable for their results in attainment and performance tables in England are based on attainment. This thesis accepts the importance of literacy and mathematical skills. Therefore, it also evaluates the P4C impact on attainment.

It becomes apparent that this study synthesised evidence from published studies, the results of a study with a comparison group, P4C training and classroom observations and analysis of secondary data from the Department for Education in order to achieve a multi-dimensional and holistic evaluation of this school-based intervention.

1.4. Thesis Outline

This thesis consists of 15 chapters. Chapter 1 is the introduction. Chapter 2 introduces and evaluates the characteristics of P4C programme. Since I am not a programme advocate, but a programme evaluator, I present the programme in a multi-dimensional way even in this theoretical chapter. Thus, I describe the programme characteristics according to Lipman's model. However, I also present a different model for philosophical discussion in the classroom, or discuss the dialogue which emerged between adherents of the programme and its opponents. I also include my personal criticism about the implementation of sessions.

Chapter 3 discusses the two main concepts: critical thinking and creativity. This thesis will provide a multi-dimensional evaluation of the programme, but it has a focus on these two concepts. Since they appear to be broad, it is crucial they are

operationalised here. Thus, Chapter 3 discusses definitions of the two concepts and concludes with working definitions for the research. Moreover, this thesis perceives these concepts as subject-independent. Therefore, this theoretical chapter also shows the scholarly debate about whether these skills are independent of a subject or are expressed within a subject.

Chapters 4-6 present the methods of this research. Particularly, the research design to address each of the four research questions is presented. Chapter 4 presents the research design of the systematic literature review. Chapter 5 presents the methods of the quasi-experimental study conducted to respond to the second and third research questions. Chapter 6 discusses the secondary data analysis, which was the method used to investigate the fourth research question of this thesis.

Concerning the second and third research question, the assessments tools used were designed for the purposes of this research and the data collection. Chapter 7 presents the measurement tools. Chapter 8 describes the marking process. Chapter 9 describes the pilot study of the measurement tools to achieve their validation for this study. I demonstrate how my pilot study satisfied the conditions to be considered successful and allowed me to proceed with the study.

Chapters 10-12 present and discuss the research results. Specifically, in Chapter 10 there is a presentation of the results of the systematic literature review. Chapter 10 examines fully the published evidence on the programme effectiveness and as a result it also demonstrates the literature gaps. Through these literature gaps, the reasons why the other three research questions (impact on critical thinking, creativity and attainment) were examined become apparent. This is a conscious decision I took as a researcher. My first research question systematically investigated the current evidence and identified potential literature gaps. I believe that the demonstration of the literature gap is not always made in a persuasive way. Therefore, I decided to conduct a literature review examining all the experimental and quasi-experimental studies conducted to evaluate P4C. Therefore, the literature gaps are discovered not only by referring to the content of the existing literature, but by calculating and examining the effect sizes of the published studies.

Chapter 11 discusses the results of the trial for the P4C impact on critical thinking and creativity. Finally, Chapter 12 includes the secondary data analysis which examines the P4C impact on attainment.

Chapter 13 presents the limitations for the research design of each of the four questions. These limitations lead to Chapter 14, which offers recommendations for future research. The final practical conclusions of this thesis can be found in Chapter 15. The available evidence is summarised and there is an overall evaluation of the programme. This chapter answers the main research question I addressed for the conduct of this thesis: does the available evidence suggest that P4C is effective and worth being implemented in schools in England? How should the programme be implemented? Can P4C develop students' cognitive and non-cognitive skills? If yes, on which domains should the educators expect change when they choose to implement this programme in their schools?

2. The intervention: Philosophy for Children (P4C)

Since this thesis deals with 'Philosophy for Children' (P4C), this chapter provides some background about the programme. Ventista and Paparoussi (2016) also discussed its implementation and suggested ways that the programme can be introduced in the classroom.

This chapter critically analyses fundamental characteristics of the programme, such as the development of a community of enquiry, the session structure and the role of the teacher in the classroom. This chapter also refers to different models or views.

This chapter negotiates both theoretical perspectives regarding the P4C programme and research evidence. The first P4C project implemented by Lipman is discussed. Then, the results of the recent meta-analyses on the topic are presented. Finally, P4C studies which examine the extent that P4C sessions are enjoyable for the students are criticised.

2.1. Community of Enquiry

P4C was developed in the USA in the late 1960s by Matthew Lipman. While Lipman was teaching Philosophy in Columbia University, he observed that his students lacked basic reasoning skills. He felt it was already too late to develop these skills at the university level and he concluded that formal logic can and should be taught at an earlier stage (Lipman, 1976; 1982). Specifically, he argued that young children are lively and curious as they begin their formal education in kindergarten (Lipman, 1976; 1982; 2003). In other words, children have a natural curiosity. When children enter the educational system, their natural curiosity and imagination seem to decline.

Lipman (2003) claimed that the environment in schools is regular and stable and it demands disciplined students who obey rules efficiently rather than independent thinkers. The educational system fails to preserve and develop the traits with which the students enter the school. Their perseverance and development could be achieved by an environment which is challenging and constantly stimulates speech and thought (Lipman, 2003, p. 13) and instructional material which arouses intellectual surprise (Lipman, 1976; 1982). Therefore, he developed the P4C programme and resources.

Primary school philosophy is about providing children with the opportunity to explore fundamental aspects of their experiences which are already meaningful to them, to become more sensitive to their philosophical dimensions (ethical, logical,

metaphysical and epistemological). P4C involves the engagement of students in a philosophical Community of Enquiry. Lipman (2003, p.20) mentioned that the term 'Community of Inquiry' was initially used probably by Charles Sanders Peirce. Peirce referred only to a scientific community, whilst Lipman prioritised the dialogical character of enquiry and the primacy of questioning. It is also about developing the ability to question, formulate an argument, wonder about things that are taken for granted, being receptive and open to the idea of others, and working collaboratively.

P4C is a movement that promotes a forum for discussions in which children are encouraged to think and reflect together, to justify their beliefs and ideas, and to become aware of their capacity for dialogue. The students are also encouraged to ask questions, because as Lipman noted: "[…] In any event, this recognition of the elevated status of the question (and the reduced status of the answer) will help the students remember that questioning is the leading edge of inquiry; it opens the door to dialogue, to self-criticism, and to self-correction" (Lipman, 2009, p. 32).

Those who participate in this community learn together and share experiences. The central aim of P4C is to help children develop their thinking for themselves and their thinking in a community. This means that even though the students think in a community where each opinion is respected, each student is not obliged to obey or conform to the opinions of others.

Lipman (2003), influenced by Dewey, discussed the idea of reflective thinking. According to Dewey (1933), reflective thinking has a purpose and aims to reach a conclusion. Reflective thinking involves 'the voluntary effort to establish belief upon a firm basis of evidence and rationality' (Dewey, 1933 p.9). In other words, Dewey suggested that reflective thinking involves inquiry of beliefs which are taken for granted. He also argued that reflective thinking is not merely a sequence of ideas but a 'con-sequence'.

According to Lipman (2003), this type of thinking does not only focus on the subject itself, but also on the procedures and methodology. For Lipman (2003, p. 27) successful thinking involves critical, creative and caring aspects combined with reflection on its own procedures. As a result, he believed that P4C encourages students to think critically, creatively, and caringly.

In this community the teacher has the role of a facilitator. Wartenberg (2009, p.8) in his book *Big Ideas for Little Kids* argued; 'You don't have to know any philosophy to teach it!' Maybe claiming that 'any' philosophy is enough to teach P4C

is problematic, but I also argue educators do not have to be philosophers to teach philosophy in primary schools. According to Wartenberg's (2009) lesson plans, the dialogue is guided by the comments of students, but the teacher is prepared and has pre-decided some leading questions deriving from the main topic of the material presented.

Lipman (1985) argued that the teachers should not impose their views but intervene only when the dialogue is turned into an exchange of anecdotes, to introduce an activity with a purpose, such as an assumption identification activity. Kennedy (2004) suggested that the teacher as a facilitator cannot pre-decide or control where the dialogue will lead. He also discussed some similarities between the role of the teacher proposed by Freire and the role of facilitator in a community of enquiry. The teacher as a facilitator summarises statements and helps the students to discuss the consequences and the assumptions of their statements (Kennedy, 2004, p. 758). Moreover, the facilitators of dialogue in a community of enquiry should always understand that each time they intervene the dialogue is slightly transformed (Kennedy, 2004: 761). If the intervention of the facilitator has an impact on the dialogue, then it is advisable that the extent to which educators intervene should not be pre-decided. It should be dependent on the needs of the specific classroom and context. There are situations which might require more guidance from the teacher. For example, when students are very young or not trained in P4C or when the topic discussed is too sensitive, facilitators may intervene more often.

As a result of this perception of the role of the teacher as a facilitator, the term 'Philosophy with Children' (rather than "for") was introduced by some P4C adherents (Sutcliffe, 2017). According to Vansieleghem and Kennedy (2011, p.178), the adjustment is important since the new term emphasises dialogue not as 'coaching', but as a generation of 'communal reflection, contemplation and communication'. The community is presented as built with the students and not as created for the students. The preposition 'with' also demonstrates that the facilitator and the children, who are the participants in the community of enquiry, are equal.

By reading literature about 'Philosophy for Children (P4C)' and 'Philosophy with Children', I do not think that there is any substantial difference between the scholars accepting each of these terms. This is due to the fact that scholars who accept the term 'P4C' do not imply any superiority of the teacher in the community of enquiry in relation to the children.

2.2. The structure of a P4C session

Having presented the main principles of the programme, this section of the chapter considers the concept of a lesson plan for a P4C session. P4C sessions are built on and with the experience of the students and from this perspective P4C sessions could be perceived as including several elements of the theory of constructivism (Golding, 2007). A simple P4C session has three main stages.

- **Stimulate the dialogue**

The facilitator of the dialogue provides the Community of Enquiry with a stimulus. The stimulus leads students to set questions and then conduct dialogue. Lipman presumed that material which stimulates the dialogue should be constructed specifically for this purpose. Therefore, he authored novels to encourage discussion and teach thinking. When Lipman (1992) discussed his first novel *Harry Stottlemeier*, he mentioned that what distinguished each character in the book from another was their style of thinking.

Lipman did not write novels just to stimulate dialogue. He explained that he wrote novels with children as characters, because he did not want to present adults as those who hold the knowledge and children as the naive (Lipman, 1992). He used characters with a similar age of the students in the classroom. These could help the students to identify themselves with the book characters. He also explained that the characters in the novels are used as a model of how a community of enquiry operates, since they are presented to dialogue.

However, the novels of Lipman are not the only possible stimulus for dialogue. P4C adherents after Lipman suggested different types of stimuli. Literature in general can be viewed as being central to the philosophical community's discussions in the sense that asking questions is a spontaneous response to literary texts that offer the reader thoughts to reflect on, new perspectives to consider, and assumptions to verify. Literary texts explore issues that matter to us as human beings, and as they present 'gaps' and indeterminacies that offer the opportunity for discussions of multiple alternative interpretations. Some P4C adherents uphold the opinion that picturebooks can successfully stimulate thinking and dialogue (Haynes & Murris, 2012; Murris,

1992). Progressively, a more general approach has also been proposed. Anything can potentially stimulate a discussion. Fisher (2003, p.111) suggests stories, poetry, photos, music, videos, or even objects as introductory stimuli in P4C sessions.

- **Ask questions and decide on the topic for discussion**.

After presenting the stimulus, the students set questions and then vote on the question(s) to be discussed in that session (Fisher, 2003). The facilitator writes down their questions – usually on the whiteboard. After the proposed questions are collected, the students decide the question that they prefer to discuss. This decision is taken as democratically as possible. The standard practice is voting through raising hands (often with eyes closed to prevent 'herd' mentality). This practice resembles the ancient Athenian democracy.

- **Dialogue**

The main part of the session involves students' dialogue. Ventista and Paparoussi (2016) suggested some indicators to assess the engagement of students in the dialogue. These indicators can evaluate whether P4C is practiced successfully in the classroom.

- *Take part in the dialogue*. For a thriving dialogue in a community of enquiry, all the students express their ideas. However, 'philosophical discussions [...] are not opinion surveys'. (Mascitelli-Morey, 2013, p.74). In the dialogic process, the students should listen to the opinions expressed by their classmates. Fisher (2005) argued that during P4C sessions disconnected answers are sometimes provided due to the eagerness of the students to express their viewpoints. The phrases 'I agree' or 'I disagree' linking answers with what the previous student has just said can be used as indicators of apparent listening.
- *Justify*. Justification is a form of reasoning which can be used by children (Thomas, 1992, p.98). The facilitator can motivate them to provide solid reasoning to support their opinion. Students should justify their opinions and provide reasons and examples from their experience. A small-scale experimental study (Gasparatou & Kampeza, 2012) with a control group (15

students in each group) in a Greek kindergarten explored the discourse and words that the students used in the sessions. The 'because' was the marker found to be used the most from the experimental group. This suggests adoption and expansive use of justification in the P4C sessions with young students.

- *Ask questions.* Fisher (2005) categorised the questions that children ask in five categories: questions that focus attention, force comparison, seek clarification, invite further enquiry and seek reasons or explanation. This approach can be important for the classroom practice. Weber and Wolf (2017) recognised the central role of question in the Community of Philosophical Enquiry and they examined types of questions and the method of questioning in P4C community. They challenged the existing approach of separating the questions to open and closed questions, with the first category to be the only category which is useful for the dialogue in the classroom. They argued that an important element in the community is to discard the unequal relationship of power between the people who dialogue. No questioner in the community of enquiry should be presented as the knower of the answer or appear in an inferior position. They explained that there is no specific method of effective questioning to be taught, even though questioning can be role-modelled. Effective questioning can lie on an attitude of openness to the unfamiliar experiences and of readiness to depart from the current beliefs and knowledge if this is necessary.
- *Define the main concept and search for criteria.* Two ways to turn a simple discussion into a philosophical discussion are defining concepts and setting objective criteria (Bassiri & Vaidya, 2013). Philosophers search for definitions and therefore students in Community of Enquiry can dialogue on questions which refer to definitions of concepts, such as 'What it means to be courageous'? Lipman (2003) argued that critical thinking employs criteria and can be assessed by appeal to criteria. Students might reason but they might also dialogue on what Lipman called 'meta-criteria' and 'mega-criteria', which will be discussed in Chapter 3.

2.3. Criticism of the structure of P4C session

After presenting the P4C structure, some of the elements will be argued as out-of-date and needing to be adjusted according to the current pedagogy. What Lipman suggested as P4C has to be in accordance with modern pedagogical methods and techniques. P4C

should take into consideration updates of educational pedagogy and research. P4C should not take place in a traditional and authoritative context. There are two main examples that I identified in order to demonstrate the need for updating some elements of P4C practice.

- **Setting Rules in a Community of Enquiry**

When a class is introduced in P4C, Wartenberg (2009, p.41) suggested that a list of rules should be posted during the first session. Wartenberg described this process as an announcement of the rules of a P4C session decided by the teacher and communicated to the students. More than fifty years after Summerhill (Neil, 1960), where the rules were decided by students and teachers, it would be odd for the teachers to establish the rules and impose them on their students.

Fisher (2003, p. 62) also referred to community rules and particularly suggested that 'these can be established by the discussion leader or agreed through discussion by the group'. He suggested though that 'whatever rules are adopted the chances of them being followed will be much greater if the children themselves have been involved in the formulation' (Fisher, 2005, p. 138). It could be argued that the set of community principles seems to be a contradictory part of a community of enquiry. P4C is a democratic intervention, whilst classroom rules have been suggested as a way to establish the authority of the teacher and help classroom management (Boostrom, 1991).

What is more, when Fisher opted to give an example of rules voted by the students he mentioned the following rule 'Don't say anything mean, stupid or unpleasant' (Fisher, 2003, p.62). It could be argued that this rule implies an authoritative context for a community of enquiry. A rule which suggests not citing anything which would offend at least one of the classmates would be a rational principle. This principle should apply in any lesson, not only P4C sessions. With reference to the 'stupid' and 'unpleasant' prohibition mentioned above, I argue that it radically contradicts the nature of the dialogue and freedom of expression. During P4C sessions, students should be encouraged to express their opinions and their thoughts unabashedly without feeling that they will be perceived as stupid. Similarly, they should be encouraged to disagree which may not be always pleasant, but the students should feel comfortable to express themselves freely. This indeed entails that they will

also learn to handle in a gentle way any disagreement which might occur in the classroom.

- **Students raise their hands to express their opinions**

Having discussed the setting of rules, which take place only during the first P4C session, the way students express their opinions should be examined. Fisher (2005, p.137) suggested four steps for this process: a) teacher or the leader of the discussion (if the leader is not the classroom teacher) asks a question b) some students raise their hands c) the leader picks somebody to talk d) while the selected person is talking, the rest of the students remain with their hands risen. This process seems to be problematic for Fisher due to the fact that the students, who remain with the hands raised, anticipate sharing their opinion, and they do not pay attention to the speaking person. This results in disconnected replies. In other words, students waiting to talk usually do not adjust their reply according to the opinion that was lately expressed. Instead, they express the opinion for which they initially raised their hands. After describing this process, Fisher (2005) attempted to suggest a solution by using an enhanced community rule. Therefore, he proposed finding effective rules to solve this problem; "The general admonition 'Everyone must listen' is not as effective as a particular rule, such as 'no hand up while someone is speaking". (p. 137).

My view is only partially in unison with Fisher on this. The described process is indeed problematic. Several students waiting by having their hands raised in order to express their opinions while one of their classmates is talking is not appropriate for a Community of Enquiry. The solution, however, cannot be an improved classroom rule. Nowadays alternative ways of talking in the school classroom are suggested and abandoning the tradition of raising hands is proposed (Brooks & Dixon, 2013). Therefore, students should attempt to self-regulate the dialogue to some extent. In a P4C session raising hands should not be a common practice, except perhaps for voting. Another technique which is currently used in P4C sessions is the palms out technique, where the students hold out their palms when they want to contribute in the dialogue. However, I find this technique similar to hands up technique.

Instead of these techniques, I suggest students initially to be trained in dialogue skills by discussing in smaller groups. Only in intense dialogue moments should the

facilitator of the discussion be responsible for choosing a speaker. If students are trained to talk freely and regulate who is speaking, then the dialogue should occur more naturally, and the replies will be more connected to each other.

- **The Learning Environment**

Learning does not take place only in the classroom. P4C should not be linked to a specific space and be presented only as a classroom-based intervention. P4C can be implemented in a less traditional setting. Last year, whilst I was participating in a research team, we investigated implementing P4C in a museum context (Ioannou, Georgiou & Ventista, 2017). Vansieleghem (2011) engaged students in Cambodia in a P4C session where the main activity was walking. These studies can also demonstrate that students do not have to be in a disciplined environment with raised hands in order to participate in a P4C session.

2.4. P4C in the UK

Since this thesis presents research conducted in England, it is important to examine a brief history of P4C in the UK. The Society for the Advancement of Philosophical Enquiry and Reflection in Education (SAPERE) was the first P4C organisation in the UK. It was founded in 1992 (SAPERE, 2015b) due to the expressed interest aroused by the documentary 'Socrates for six years old' which was on BBC in 1990 (Sutcliffe, 2017). SAPERE considers Lipman's approach the 'gold standard' (Sutcliffe, 2017, p. 5).

Currently, there are different organisations offering P4C training in the UK. However, in this thesis there is a specific reference to SAPERE for two reasons. First of all, it was the first P4C organisation established, and therefore, it has the longest tradition in the country. Secondly, SAPERE assisted with the comparative evaluation study conducted in this thesis. Particularly, empirical data was collected in order to reply to the second and third research question. This involved schools which were engaged in implementing P4C sessions. For consistency reasons, it was crucial for the schools to have received the same training. Therefore, SAPERE kindly supported me with this project and helped in the recruitment of P4C schools.

As Sutcliffe (2017), a founder member of SAPERE and Chair for several years, mentioned, SAPERE supported P4C in the UK by offering a sound rationale for its implementation. I consider their support in my project as a continuation of the effort of

SAPERE to investigate evidence concerning the effectiveness of the programme. I am grateful for this support and I also recognise that SAPERE's search for sound evidence about P4C implementation is compatible to my quest for evidence-based educational policy.

2.5. Different models

So far, this chapter discussed the P4C model of Matthew Lipman, which is also the model that SAPERE follows. This model is the focus of this thesis because the teachers of the intervention schools participating in the research were trained by SAPERE trainers, who promote this model. However, other main models for doing philosophy with children should be briefly discussed. Despite the common idea of philosophising with children, the different versions have fundamental differences and it might be questionable to what extent they can be considered as a united programme.

Nelson's Socratic method is not going to be discussed by this thesis, because it is a model used for adults. However, it is essential to refer to the McCall's model of 'Community of Philosophical Inquiry'. The similarities and the differences between the two models are discussed by McCall (2009) who developed this model several years after Lipman's model and the foundation of SAPERE. Table 2.1. summarises the comparison between the two.

Table 2.1. Comparison of different Community of Enquiry models.

Elements	P4C model (Lipman and SAPERE)	Community of Philosophical Inquiry (McCall)
Philosophical theory underlying	Pragmatism and the ideas of John Dewey	Realist Philosophy
Introductory stimuli	Lipman's novels or according to SAPERE various stimuli	No manuals. The Chair of the discussion plays a crucial role to structure the session.
Question chosen	SAPERE suggests voting on the question to be discussed	The CoPI Chair chooses a productive question for discussion
Teacher	Facilitator in the discussion and without philosophical background	Leader in the discussion and with philosophical training
Training of the leader of the discussion	Not compulsory for the teacher to have training in	Compulsory for the teacher to have philosophy

	philosophy	training (minimum 80 hours) by a somebody who holds a PhD in Philosophy
Opinion of the participants	The participants usually express the opinion.	The participants are explicitly told that they can contribute an idea for the progress of discussion and this does not have to be their personal opinion.
Nature of the discussion	Open	Structured Reasoning

The model followed by Lipman and SAPERE appears to be more liberal and pedagogical compared to McCall's. In the latter, the Chair of the discussion chooses the question. The democratic element missing from the CoPI approach is a significant element in P4C, and it can be linked with the development of caring thinking that Lipman aspired to. Recent interviews with students who participated in P4C session revealed P4C as a participatory pedagogy where the students are free to express their opinion and regulate the talk (Barrow, 2015). This important element is missing from McCall's model. Moreover, the training of the Chair person is quite extensive in the CoPI and it might discourage several teachers introducing P4C in their classroom.

2.6. Philosophy for Children and Developmentalism

This thesis is not a unidimensional presentation of P4C. Hence, this chapter discusses oppositional views, such as those presented by some developmental psychologists. The P4C movement is in agreement with a specific narrative of childhood. An effective implementation of P4C requires the teacher to see the child as an agent and competent, passing control to the child and this narrative of childhood is distinctively different by the way developmental psychologists perceive children.

Developmental psychologists perceive children as 'developing' and childhood as 'becoming' instead of 'being' (Lyle, 2017). Therefore, children are expected to be able to complete specific cognitive tasks at a specific age. This is also compatible with the idea that children should only ever be taught in the school in a way that it is age appropriate. Hence, the focus of this section is the refutation of the arguments that developmental psychologists made.

This contradiction between P4C and developmentalism with Piaget being its main advocate has been discussed extensively by Gareth Matthews (1978; 1984; 1994)

when P4C was still a new movement. If the developmental stages of Piaget were accepted, that would mean that P4C cannot be effectively taught. Matthews (1978) highlighted that the results of Piaget experiments establish what the majority of the children can do at a specific age. The way the experiments are conducted and the aim of establishing developmental stages based on the majority of the students results in exclusion of unique or rarely repeated responses. The philosophical puzzlement which might occasionally occur is not incorporated in the developmental stages of Piaget.

Matthews recognised the characteristics which make the theory of Piaget attractive (Matthews, 1994): arresting results, which are replicable and demonstrate an age-related sequence. However, Matthews explained that thinking should not be perceived developmentally in the same way that walking would be. Children might be developmentally unready for walking, but they are not developmentally immature for thinking. Matthews provided his readers with a lot of examples were young children naturally philosophise (Matthews, 1984).

Moreover, Matthews identified a significant flaw in the theory of Piaget. By accepting the developmental stages, Piaget discussed the transition of less to more sophisticated replies. This notion includes an assumption implying that the adults are able to provide more sophisticated replies than children (Matthews, 1978). In this sense, as Lyle (2017) argued, childhood is perceived as a preparation for adulthood. Taking this assumption for granted is erroneous according to Matthews. Matthews counter-argued the developmental stages of Piaget and the transition to more sophisticated answers by presenting two different examples. In the book *Growing Up with Philosophy*, Matthews discussed the developmental stages concerning the development of thought, while later he devoted a complete chapter for the development of thinking concerning conservation in his book *Philosophy of Childhood*. At this point, I will introduce and evaluate his arguments. I consider these arguments particularly important, because they can be considered a defense of P4C movement in opposition to developmentalism.

Concerning the question ‘what do you think with’ set by Piaget (Matthews, 1978) his developmental stages suggest that the children at stage one believe that they think with the mouth, at the stage two (age 8) they believe that they think with the brain, while at stage three (age 11-12) their thought is no longer materialised. Matthews matched each of these responses with philosophical theories: stage one response with Plato’s inner speech and the writing of behaviourist Watson, stage two

with identity theory and the idea that mental events are identical to brain events and stage three with classical dualistic theories. By suggesting this parallel between developmental stages and philosophical theories, Matthews' argument tries to show that there is no naive or less sophisticated reply. On the contrary, each response is sophisticated, and it demonstrates a philosophical puzzlement. It might also be the case that younger children are more puzzled and as they grow up they tend to conform by adopting responses accepted by society.

Similarly, Matthews (1994) discussed the association between different child responses concerning conservation of substance, weight and volume at different ages and philosophical theories. For example, he associated the egocentrically related responses at stage one with the theory of Protagoras, according to which "man is the measure of all things" (p.48). In this argument of Matthews the egocentric perception of children at this stage does not appear as naive.

However, the second argument of Matthews is not equally convincing. Matthews defined extreme and moderate egocentrism, as a lack of interest or failure of a person to recognise the feelings of other and demonstrated that Piaget did not examine whether the young children imagine how things (and specifically the clay in the conservation experiment) feel. Matthews believed that this is how egocentrism should be defined and that the experiments of Piaget did not capture this element. However, Piaget did not define egocentrism in the same way. Piaget referred to egocentrism of the child as the phase that 'the things are considered to depend solely on his personal activity' (Piaget, 1999, p. 366). This has many implications. For example, during that stage space is perceived as a function of the child's own body instead of locating the body in space (Piaget, 1999, p. 204). The experiments of Piaget were consistent with the definition of egocentrism in his theory and therefore introducing a different definition and accusing Piaget of not measuring the elements of the new definition is not a convincing critique of his theory.

Matthews concluded his chapter with what I perceive as his strongest argument critique of Piaget's conservation experiment. He discussed whether the conservation of substance, weight and volume can be questioned. It is energy and mass which is conserved at the end, not substance. The weight varies if we are not in the Earth and finally even the example of volume conservation can be questioned. By saying this, Matthews makes a clear point and questions whether it is valid for the stage during

which students accept ambiguous statements to be accepted as the ultimate stage of cognitive development.

To sum up, Matthews developed a critique of developmental theory. He demonstrated in his work different examples of philosophical puzzlement in children's discourse. This discourse is not repeatable, and the theory of Piaget, which is mostly normative, fails to capture it. The assumption of moving from less to more sophisticated cognitive stages fails by considering that the adults are not always reasoning and holding correct beliefs and by considering that the replies of students during Piaget experiments can be associated with different philosophical theories.

Lipman (1976; 1982; 1987) also made references to Piaget's work and his developmental stages which support that students in primary school are restricted only to concrete reasoning and experience. He suggested that developmental psychologists focus only on what students can do without intervention.

To summarise, the contradiction between developmental stages and P4C is an important theoretical debate, since the acceptance of the first might lead to questioning the effectiveness of the second. Nevertheless, Piaget's theory can only be correct when there is no intervention. What Piaget considers as less sophisticated or wrong responses can be questioned. The fact that children usually provide specific answers at specific ages does not exclude the possibility of changing these stages by using intervention programmes, such as P4C.

2.7. The first Philosophy for Children project

Having presented theoretical elements of P4C, the implementation and evaluation of the programme should be considered. The first P4C project was conducted by Lipman in the fall of 1970. Lipman presented the results of the project (1976;1982) which he conducted. As he said, he thought that reasoning should be taught in a more systematic way in childhood and in 1969 he applied for funding to the National Endowment for Humanities for a pilot project grant proposing to write his first P4C manual and conduct a pilot study. When the study was approved, he conducted a study with an experimental design. Each group had twenty children and the children were randomly allocated to the groups. The project lasted for nine weeks. The intervention group was taught P4C twice per week by Lipman himself and two teaching assistants. The comparison group received science. Initially, for the first three weeks the comparison

group received an alternative treatment, but after three weeks this collapsed. Thus, they received regular social science instruction instead of an alternative intervention.

Before presenting the results of Lipman's study, I will evaluate this research design. First of all, the sample is too small and even though Lipman claimed randomisation, when the sample is too small then randomisation cannot be considered effective (Gorard, 2013, p.128). The project lasted for a short period which seems difficult to cause actual impact. The most unrealistic part of the project is the teacher-students ratio. Lipman had two teaching assistants for a class consisting of 20 students. This means that each teacher had responsibility for approximately seven students. This is very uncommon for classrooms in state-funded schools in the UK. Therefore, there are many weaknesses in this research design, since it makes the study unrealistic and it cannot lead to generalisation.

Concerning the results, Lipman (1976;1982) admitted that he initially thought that the programme had no impact because of a computer printout received which claimed lack of statistical significance testing in the reasoning tests. However, he claimed that in summer of 1973 he read a report by one of the teaching assistants of the project which claimed that with initial equivalence of the two groups, the experimental group showed statistically significant results in logical reasoning. This seems a bit bizarre. It becomes even stranger when Lipman, after having read the report, attempted to verify this result by calling Jerry who was the teaching assistant who wrote the report. As he said

> It took me several days to digest this information. How significant was the reported difference of .01? [...] I could hardly believe we'd made such an impact on the kids in the study. After all, we'd made much of a fuss about teaching logic: there was no homework, no grades, no written classwork [...] I called Jerry. He told me that the results were quite as he had set them down in his report. Unfortunately, he no longer had the data, which meant that our finding couldn't be substantiated.

This indeed seems a quite peculiar conclusion for a project. Therefore, the impact of the first P4C project on reasoning skills, according to the person who set up the project, could only be verified by a report written by Jerry[1]. Even though Lipman set up the project, he could not find the data when he read the report.

[1] He refers to Jerry Jaffe.

Consequently, I conclude that Lipman project did not show any evidence for the impact it has on reasoning. Lipman attempted to investigate this impact further by comparing the reading scores of the two groups. This comparison was presented in the Bierman report (Lipman, 1976; 1982). The report claimed significant impact on reading skills for the experimental group. Nevertheless, there is no evidence that the two groups two and a half years later were still equivalent. Additionally, as it is reported (Lipman 1976; 1982), even though all the participants of the experimental group were retained, five students were missing from the comparison group. It is evident that this is a significant amount of missing data given that initially there were only twenty students in the comparison group.

To summarise, the first P4C project had weak research design and did not provide any generalisable results. This study led to the publication of two reports concerning the impact of the programme on reasoning skills and reading. None of the two reports is robust. The one report was untrustworthy because even Lipman who set up the project could not obtain the data to confirm the programme effectiveness. The second report occurred two and a half years later with a high proportion of missing data.

2.8. Evidence about the P4C effectiveness

Although the first P4C project reports are considered untrustworthy, ensuing evidence sheds some light on the effectiveness of the programme. Trickey and Topping (2004) conducted a systematic literature review and they included ten studies of which eight reported a positive impact of the programme on different cognitive and non-cognitive skills and two provided insufficient data for conclusions to be drawn. However, this systematic literature review was published in 2004 including the studies published until 2002, more than fifteen years ago. However, curriculum reform might take place approximately every ten years (Sargent et al., 2010). Consequently, it is likely that education reforms might have occurred and this positive impact might not still be applicable in recent studies.

The same year a meta-analysis which evaluated P4C was published by Garcia-Moriyón, Rebollo and Colom (2004). The meta-analysis focused on studies which evaluated the impact on reasoning skills and included 18 studies. The authors recognised that 17 of the studies reported positive effect sizes. However, they also recognised that there might be publication bias and studies reporting positive results of

a programme are more likely to be published. This means that they did cover a holistic and multi-dimensional examination of the topic. Additionally, the meta-analysis by Garcia-Moriyón et al. (2004) focused on studies which examined the impact of P4C only on reasoning skills. Since 2004, however, many studies have been published which examine the effectiveness of the programme.

There is a recent meta-analysis by Yan (2017), but it is currently under an embargo. I could only access its abstract. The abstract of the study suggested that studies published from 2002 to 2016 show medium effect sizes for the programme in different areas, and big positive effect size for reasoning skills.

There are some studies examining to what extent P4C is a joyful experience for students and teachers. Some examples are worth analysing since they have some similar characteristics. In general, the students appear to provide positive feedback concerning their participation in P4C sessions. Research in Northern Ireland investigated the perceptions of 364 students and 19 teachers who participated in a sub-category of P4C and suggests that students enjoyed participating in P4C sessions (Dunlop, Compton, Clarke & McKelvey-Martin, 2015). It is worth noting that the interviews conducted with 16 teenagers in Greece after P4C sessions also assigned positive feedback (Gasparatou & Ergazaki, 2015). When students were asked what they enjoyed more, they mentioned the lack of the demand for providing right answers. P4C sessions are based on the notion that there is no right or wrong answer. This is one of the central beliefs that educators and students have about P4C sessions, but it is not warranted. 'Some answers are simply and plainly wrong, some are better than others' (Gazzard, 2012, p. 52). Learning this is part of what philosophy is.

What is more, these studies were sometimes weakly designed. What I found intriguing in one of these studies (Reznitskaya & Glina, 2013) was the fact that the researchers decided to examine the opinions of participants in P4C by conducting an experimental trial and having a comparison group. There was also randomisation within the groups. It is odd that the researchers chose this experimental research design since this design does not fit the research question. It seems that in the interviews the researchers included questions highly-related to the content of the intervention. For instance, the students who participated in P4C sessions mentioned that they liked the disagreement during dialogue. It is not surprising that this is less commonly mentioned by students with regular classes, as it is a basic element of a P4C intervention.

To sum up, this category of studies remains only on a superficial level of whether P4C can 'entertain'. Consequently, there is no apparent reason why this research question is repeatedly examined and why there are so many replications studies examining the same research question. The problematic element is that several of these studies involve just a few sessions and students' interest and engagement are expectedly retained. I assert that a study, which asks the participants whether they enjoyed an intervention and whether they felt engaged, should adopt a longitudinal approach covering a considerable amount of P4C sessions to verify whether engagement and interest are retained. The studies which examine whether the students enjoy P4C sessions ask the participants to be engaged in sessions and teachers to offer didactic time and energy for their implementation.

I expect that educational research should offer informative results for the public and the policy. I argue that replicating studies with this research design to examine the enjoyment of the students does not respect the time of their participants. This is due to the fact that there are already too many studies suggesting that short P4C interventions are enjoyable. I consider it a loss of time to replicate more studies to verify that the students like a change in their routine by participating in a few P4C sessions. For future studies interested in finding out how enjoyable P4C sessions are for the students, I recommended to scrutinise the interest in sessions with a longitudinal design following the same cohort for years.

2.9. Chapter Summary

Since this thesis evaluates P4C programme, this chapter presented the characteristics of the programme. Initially, it presented the concept of the Community of Enquiry and the structure of a P4C session based on Lipman's and other P4C adherents' writing. However, for a balanced presentation of the topic, criticism on P4C was also discussed. For this reason, I criticised some elements of the P4C structure explaining why the structure of a session urges updating. Then, oppositional views were discussed and refuted. Finally, P4C research projects and evidence were critically analysed. The presentation of evidence demonstrated that there is no recent combination and evaluation of the evidence to inform evidence-based decisions for educational policy in the UK or internationally.

3. Defining the constructs: Critical Thinking and Creativity

This thesis examines the impact of P4C on cognitive and non-cognitive skills. The second and third research questions focus on the impact on critical thinking and creativity respectively. The way these concepts are defined plays a crucial role in the way they are assessed.

Given that the impact on critical thinking and creativity had to be measured, it was important to define these concepts. For these two concepts this thesis uses the term 'construct', which is defined as the concept or characteristic that a test is designed to be measured (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Constructs cannot be directly observed. Therefore, for critical thinking and creativity to be measured, they have to be 'operationalised' - a term used by Stevens (1935). Therefore, this chapter presents the definitions of the constructs and concludes with the working definitions of creativity and critical thinking used in this research. Their working definitions are used to decide on and develop their measurement tools, which should encapsulate the essential aspects of critical thinking and creativity.

Initially, this chapter is going to present theoretical definitions of the constructs of critical thinking and creativity. There is still no consensus on a definition for these constructs. Therefore, this chapter discusses diverse definitions of the constructs. Also, there has been long discussion whether these two constructs should be perceived, taught and assessed as general skills independently of a school subject or within a specific domain or a school subject, and indeed whether they can be taught at all. Finally, this chapter discusses whether these two skills are malleable.

3.1. Critical thinking: Definitions

In this section, the definitions of critical thinking are presented and evaluated. These definitions lead to the debate about whether it is a general or a subject-specific skill. The presentation of definitions is important in order to formulate a working definition. The working definition of critical thinking determines the content to be assessed in critical thinking assessment used in this study. Since this thesis discusses the P4C programme, it is preferable to start discussing the definition of critical thinking that Matthew Lipman suggested. Lipman introduced his own definition of critical thinking. He suggested that critical thinking has three main characteristics:

- It is self-corrective. Lipman (1987) perceived critical thinking similar but not identical to metacognition. He claimed that 'Metacognition is intellectual self-consciousness: the mind turns on itself and thinks about its own thinking. But it can do so without thinking self-correctively. One can think about one's own thinking and yet do so uncritically' (Lipman, 1987, p.5). Therefore, critical thinking is an active exploration about and the correction of one's thinking flaws.
- It involves thinking with criteria. He argued that 'critical thinking is reliable thinking that both employs criteria and can be assessed by appeal to criteria' (Lipman, 2003, p.212). In his book *Thinking in Education*, Lipman (2003) explicitly said that criteria are reasons. Earlier in his work, he upheld the opinion that every reason presupposes a criterion, and therefore providing good reasons is linked to the quality of thinking (Lipman, 1987). He also established the importance of reasoning by introducing two main criteria for a reason to be considered good: relevance and the strength of the reason (Lipman, 1987; 2003).

  Lipman (1987, 2003) mentioned four different types of criteria; formal, informal, mega-criteria and meta-criteria. Informally anything can be a criterion in a comparison with something else. Formal criteria are those which are accepted in some institutionalised context, such as laws and regulations. Mega-criteria are generally accepted and presupposed. Meta-criteria refer to criteria, such as strength and relevance, which are criteria used for the choice amongst criteria.
- It is sensitive to context. Even though Lipman (1987) referred to formal criteria and mega-criteria, which are characterised by some generality, he also accepted that critical thinking is sensitive to context. Therefore, critical thinking recognises that there are exceptional situations and limitations. Consequently, critical thinkers would not overgeneralise or transfer from one context to another, when this is not appropriate.

According to Daniel and Auriac (2011, p.420), the definition of critical thinking by Lipman is 'pragmatic, in that for him, critical thinking is a complex process that is integrated into a utilitarian design for the improvement of personal and social

experience'. It should not be surprising that the definition of Lipman was characterised as pragmatic since Lipman (2003) was strongly influenced by the pragmatist Dewey.

Having presented the definition of critical thinking by Lipman, I argue that even though the conceptualisation of critical thinking by Lipman appears to be quite precise, some of his ideas could not be easily used for the operationalisation of critical thinking for an assessment. Self-corrective thinking is an internal process and it cannot be easily observable. In that sense, this thesis, which aims at the operationalisation of the concept of critical thinking for its assessment, could not use Lipman's definition of critical thinking. Nevertheless, the second and the third element that Lipman suggested can be a guide for critical thinking assessments, which require students to use reasoning and solve problems within a particular context and were considered for the assessments in this thesis.

There is no consensus about a critical thinking definition. Nevertheless, there are specific elements of critical thinking which seem to appear repeatedly in different definitions. Even though the discussion of critical thinking in this thesis started with the definition of Lipman, there are a few other elements to be discussed that are topics commonly emerging in critical thinking definitions. These are the ambiguous elements of the definition and anyone who attempts to define critical thinking should decide on a clear stance.

3.1.1. Critical Thinking as a Guide to Action

One of the most influential scholars in the area of critical thinking is Robert Ennis. A discussion of critical thinking could not have been complete without referencing the work of Ennis, who published in the area of critical thinking not only as a theoretician but also as an educator trying to operationalise critical thinking for teaching and assessment. Ennis developed critical thinking assessments, such as the Cornell Test of Critical Thinking (Ennis & Millman, 2005) and the Ennis and Weir (1985) essay to measure critical thinking. Consequently, Ennis perceived critical thinking as a construct which could be measured.

The first definition that Ennis (1962) introduced described critical thinking as the 'correct assessing of statements'. McPeck claimed that Ennis had a narrow perspective of critical thinking by perceiving critical thinking as the *correct assessment of statements* (McPeck, 1981) because other activities also involve critical thinking. McPeck identified the lack of explanation of whether the adjective 'correct' implies

correct answer or correct thinking procedures. In the latter case, a person might employ reasoning, but fail to be correct. Furthermore, McPeck argued that the definition of Ennis provided an absolute perception of critical thinking with somebody being right or wrong and nothing else.

Later and since, the working definition of critical thinking according to Ennis (1985; 2015a) is 'Critical thinking is reasonable reflective thinking that is focused on deciding what to believe or do'. Norris and Ennis (1989) explained each aspect of this definition. Specifically, critical thinking is reasonable because it is based on reasons, which leads to the best conclusions. It is reflective because the critical thinker also examines the reasonableness of their own thought and the thoughts of others. Critical thinkers do not merely happen to find reasons, but they consciously seek them. Critical thinking is focused which means it has a specific purpose. Finally, critical thinking has a practical element, since it assists the thinker to decide what to believe or act.

I argue that Ennis' definition is highly influenced by pragmatism and John Dewey. Dewey (1933) argued in favour of reflective thinking and explained why reflective thinking should be an educational aim. Furthermore, pragmatist Peirce emphasised on the importance of beliefs as a guide to action. Instead of doubting everything not based on sound foundations as Descartes would suggest, pragmatist Peirce suggested that beliefs should be doubted when they fail to guide action successfully (Pring, 2007).

I argue that this perception of critical thinking as a guide to action can be considered utilitarian. It is focused on the usefulness of thinking and therefore it shows the applicability of critical thinking skills in daily life and its consequences. This definition suggests the importance of thinking for life. The impact of critical thinking is observable and assessable.

Fisher and Scriven (1997) disagreed with the fact that critical thinking should be perceived as a guide to action because people might decide to act in different ways. Choosing to act irrationally does not mean that they do not possess critical thinking abilities.

3.1.2. Critical Thinking as Problem-Solving

As it has already been discussed, according to Ennis (1985; 2015a): 'Critical thinking is reasonable reflective thinking that is focused on deciding what to believe or do'. The last part of the definition implies a problem-solving element. It can be questioned

though to what extent critical thinking is equivalent to problem-solving. This aspect of critical thinking is usually included in its definitions. In some cases, problem-solving is a component of the definition, whilst in others it is the central part of the definition.

Halpern (1998) who is the author of a critical thinking assessment (Halpern critical thinking assessment)[2] defined critical thinking with five dimensions:

- verbal reasoning. So far, in all the definitions of critical thinking in this chapter, reasoning is always included.
- argument analysis skills. Real-life problems according to Halpern are complex and they do not only include statements, but also assumptions, intermediate steps and irrelevant information. Therefore, the argument analysis skills are very important.
- skills in thinking as hypothesis testing. These are useful because people are usually required to predict and explain information, generalise and test the validity of hypotheses.
- assessing degrees of likelihood and uncertainty, because only a few events can be known by certainty.
- decision-making and problem solving. One of the main purposes of using critical thinking skills is to make decisions and solve problems. In this sense, Halpern (1998) suggested that critical thinking includes judgement, generating and choosing amongst alternatives. This entails that critical thinking also includes some creative thinking, since it involves the generation of alternatives.

Consequently, it can be said that Halpern perceived problem-solving as an aspect of critical thinking. Sternberg (1986, p.3) defined critical thinking as 'the mental processes, strategies and representations people use to solve problems, make decisions, and learn new concepts'. Problem-solving is the central skill in his definition. His definition involved metacomponents, performance components and knowledge acquisition components. Metacomponents are higher order processes and involve recognising that a situation or problem exists and thinking of a strategy to solve it. Metacomponents also include the monitoring of the strategy implementation and evaluation of the situation after it is solved. Performance components are used in order to implement the metacomponents and this might involve processes such as inductive

[2] The assessment targets participants who are aged 15 years old and older. Thus, it could not have been used for the purposes of this study.

or deductive reasoning. Finally, knowledge acquisition refers to the information required to learn concepts or procedures. This entails that the person should choose relevant and useful information, synthesise and compare it to previously learnt information.

This definition has weakness and particularly Lipman (1987) criticised this definition as being focused only on problem-solving. He claimed that critical thinking does not exist only in order to solve problems and it seems that Sternberg did not consider other cases where critical thinking is demonstrated. However, Lipman emphasised the word only, which means that Lipman accepted the idea that critical thinking becomes apparent in situations where the individuals might proceed to problem-solving, but it is not restricted to these situations.

This thesis argues that problem-solving is an element of critical thinking in agreement with the ideas of Halpern (1998) and Sternberg (1986). However, problem-solving and critical thinking are not synonymous and the first is only an element of the latter. For example, people who read the newspaper might judge the credibility of the sources without having to take a decision or solve a problem. This judgement might take place prior to storing this information in their long-term memory as knowledge.

3.1.3. Distinguishing between Critical Thinking and Critical Thinker

Various definitions focused on distinguishing between the critical thinking as a skill and the critical thinker. This is due to the fact that somebody might hold a skill without using it and therefore it is important to decide whether the evaluation focuses on the skill itself or the thinker. The relationship between critical thinking abilities and the thinker is investigated.

For Siegel (1988), critical thinkers recognise the value of critical thinking. He mainly connected critical thinking with reasoning. He explicitly said that critical thinkers should be able to 'assess reasons and their ability to warrant beliefs, claims and actions properly' (p. 34). He also attempted to distinguish a critical thinker from a rational person (Siegel, 1988). Siegel did not connect critical thinkers solely to their ability to reason. He also argued that critical thinkers have the skills and the attitudes, character traits and habits of mind. He named this 'critical attitude' or 'critical spirit' (Siegel, 1988, p.39) and included characteristics such as inclination to seek the truth and not only the skill of reasoning.

The most interesting element of this critical attitude is probably the disposition of the critical thinkers to use their judgment even when this judgment contradicts their self-interest. The ideas of Siegel could also be the basis for a different type of assessment of critical thinking. It could be argued that if the views of Siegel are accepted, then critical thinking assessment would involve both assessing pieces of reasoning and the critical attitude of the person.

Ennis (1985) accepted that critical thinking involved both dispositions and abilities. Norris and Ennis (1989, p.9) agreed that the abilities are not adequate for critical thinking. People need both the abilities and the tendencies to use them. Thus, Ennis accepted the inclusion of critical thinking dispositions and he categorised them into three categories (Ennis, 1996; 2011). According to these the critical thinker should:

- Care that their beliefs are true, and that their decisions justified; that is, care to "get it right" to the extent possible, or at least care to do the best they can
- Represent a position honestly and clearly. This referred to their own positions and the position of others.
- Care about the dignity and worth of every person.

He revisited this definition in several publications and in the most revised definition of critical thinking dispositions he included the following critical thinking dispositions (Ennis, 2015; 2015b, p.32):

1. Seek and offer clear statements of the thesis or question
2. Seek and offer clear reasons, and be clear about their relationships with each other and the conclusion
3. Try to be well informed
4. Use credible sources and observations, and usually mention them
5. Take into account the total situation
6. Keep in mind the basic concern in the context
7. Be alert for alternatives
8. Be open-minded
    a. Seriously consider other points of view
    b. Withhold judgment when the evidence and reasons are insufficient
9. Take a position and change a position when the evidence and reasons are sufficient
10. Seek as much precision as the nature of the subject admits

11. Seek the truth when it makes sense to do so, and more broadly, try to "get it right" to the extent possible or feasible
12. Employ their critical thinking abilities

The 'critical spirit' (Norris & Ennis, 1989, p. 11) motivates critical thinkers to use their abilities. This is what Ennis focused more when he started discussing the nature of critical thinking. In his publication, *A Definition of Critical Thinking* in 1964 he included only nine skills as major aspects of critical thinking. These skills focus on judgments that the critical thinker should be able to do. According to Ennis a critical thinker should be able to judge whether (Ennis, 1964, p.599):

1. A statement follows from the premises.
2. Something is an assumption.
3. An observation statement is reliable.
4. A simple generalization is warranted.
5. A hypothesis is warranted.
6. A theory is warranted.
7. An argument depends on an ambiguity.
8. A statement is overvague or overspecific.
9. An alleged authority is reliable.

Later, Ennis added more skills in his critical thinking definition (Ennis, 2011; 2015a). These were summarised under five broader categories:

- basic clarification, such as focus on the question
- basis for decisions, such as judge the credibility of a source
- inference, such as deduction and value judgments
- advance clarification, such as assumption identification and define concepts
- auxiliary abilities, such as rhetorical strategies

Ennis probably observed that these categories blur. Thus, metacognition and monitoring thinking were included in advanced clarification in his latest revision (Ennis, 2015a) while earlier were judged as auxiliary abilities (Ennis, 2011). Similarly, dealing with fallacies was presented in two different categories in different revisions of the critical thinking definition (Ennis, 2011; 2015a). Nevertheless, the exact categorisation of the skills is not the most important elements of these definitions. The

most important element is the fact that Ennis operationalised the nature of critical thinking with skills that could be assessable.

Particularly, as it has been previously stated, Ennis started defining critical thinking by referring to skills instead of dispositions. Even though he discussed dispositions and he included them in his final definitions, he did not seem particularly keen on the inclusion of these dispositions in the critical thinking definition and even more in assessments of critical thinking.

In the *Palgrave Handbook of Critical Thinking in Higher Education*, Ennis (2015b, p.37) wrote about a personal communication he had with Stephen Norris. Norris accepted that dispositions might have been more important than the abilities since the abilities are important only if they are used. However, Ennis did not seem to agree with this position.

Ennis (1996) argued that the dispositions of people cannot be easily assessed. A good assessment of the dispositions would involve a one-to-one observation. This observation would be time-consuming and expensive since the observer would have to wait for an opportunity or context for the disposition to appear. Hence, Ennis presented this observation almost as an infeasible method of assessment. On the other hand, if the dispositions are evaluated with questionnaires or multiple-choice questions, Ennis (1996) reported that it is easy for the test takers to guess what the test maker would like them to answer.

Moreover, Ennis (1996) summarised some bias that might exist in critical thinking dispositions. They might involve gender bias, whilst other dispositions can be considered either good or bad. For example, the critical thinking disposition '*caring about others'* might lead to additional bias. Caring might result in unfairness and unclear judgments and this is not acceptable in critical thinking. These arguments of Ennis are persuasive.

In addition to Ennis' concerns about the assessment of critical thinking dispositions, Fisher and Scriven (1997) felt that the dispositions are not important for judging whether somebody has critical thinking abilities. Even though they mentioned dispositions, they distinguished critical thinking from the critical thinker. As a result, in their definition of critical thinking they did not include any attitudes for the critical thinker.

Consequently, critical thinking dispositions might not be assessed effectively, and it might be questionable to what extent critical thinking dispositions define critical

thinking or critical thinker. Hence, critical thinking dispositions were not used in the measurement of critical thinking by this thesis. It was not feasible to assess critical thinking dispositions in an authentic way and a self-administrative questionnaire was not judged trustworthy. Furthermore, I argue that somebody might hold critical thinking dispositions and value critical thinking without having developed critical thinking skills.

3.1.4. The relationship between critical thinking and creativity

This thesis focuses on both critical thinking and creativity. It has to be recognised that there are also definitions which combine critical and creative thinking. These definitions might use the term *critico-creative thinking* (Fisher, 2010, p.13). Critical thinking can be linked to creativity, because the first involves the imagination of alternative options which could be considered elements of the latter. In fact, when I attempted to validate established measurement tools of critical thinking and creativity correlations were found between the assessments of the two constructs, which could suggest that the two constructs are related (Ventista, 2018a). Nevertheless, in order not to overcomplicate the concepts of the thesis, this thesis examines these constructs separately. In a following section, creativity is defined separately to critical thinking.

3.1.5. Is critical thinking value-neutral?

According to the Critical Thinking movement, critical thinking can be perceived as a combination of skills. However, it can be questionable to what extent critical thinking is a set of skills which are value-neutral. If critical thinking is not used by the thinker for everyday life, then it can be considered value-neutral. However, Barnett (1997, p. 16) rejected the way that other scholars from the Critical Thinking movement perceived critical thinking. He identified two main weaknesses in their positions. Firstly, it was assumed that particular cognitive processes can be called critical thinking. Secondly, critical thinking was considered an 'assembly of skills' which are value-neutral. Barnett (1997) found this perception problematic and he talked about criticality and a curriculum of critical being.

According to his model, critical beings are not only critical when knowledge is considered. They are also critical in the domains of self and the world (Barnett, 1997, p. 103). The lowest level of criticality involved critical skills. When knowledge is considered, Barnett talked about critical reason, whilst the two domains of self and world were named critical self-reflection and critical action.

It becomes apparent that Barnett considered criticality, but combined elements of the previously discussed theories. To be more precise, he introduced four levels of criticality which separated across the three aforementioned domains. He named the lowest level 'critical skills' and in relation to the knowledge, he called them 'discipline specific critical thinking skills'. These could be linked with the theory of McPeck which is presented later in this chapter. On the other hand, when critical skills refer to the world, he referred to problem-solving skills, which could be linked to the perception of Sternberg about critical thinking.

Despite Barnett's views, this thesis accepts that critical action refers to critical thinkers and not the skills. I argue that critical thinking skills are tools which can be used in different ways. Some people might choose not to use these tools. This does not mean that they do not have them. People sometimes choose to act irrationally. I argue that an irrational action does not suggest that the person lacks critical thinking skills in the same way that an immoral action does not imply that a person is not aware of or lacks ethical values.

3.1.6. Critical thinking as an active process

Another element of critical thinking definitions is to what extent critical thinking is active thinking. Fisher and Scriven (1997, p.21) defined critical thinking as '[…] skilled and active interpretation and evaluation of observations and communications, information and argumentation'. Critical thinking is *'skilled',* and this entails that somebody might be more or less skilled in critical thinking (Fisher & Scriven, 1997; Fisher, 2010). Additionally, critical thinking is *'active'*. According to Fisher and Scriven (1997), this active element involves four levels.

The first level entails that critical thinking is not merely a passive process of comprehending, but it involves searching for equivalent meanings and identifying key ambiguities. The second level explains that critical thinking is active in the sense of being proactive. Critical thinking does not simply involve listening or reading with understanding. It involves finding further sources in order to obtain information. Fisher and Scriven (1997, p.25) also explained that empathy is an example of the proactive level, since it *involves the investigatory effort of projecting oneself into the shoes of another*. The third level is called reflective. This is the part where Fisher and Scriven (1997) suggested that reflection refers both on reflecting about the thinking of others and self-reflection, which is the same as Lipman's idea that critical thinking is a self-

reflective type of thinking. Finally, the fourth level of the active critical thinking involves formulating new principles. However, Fisher and Scriven (1997) realised that this is a complex process and therefore it is not a necessary requirement for somebody to be a critical thinker.

3.1.7. Should critical thinking be considered a general or a subject-specific skill?

Different positions have been developed for critical thinking, its nature and its definition. One of the biggest debates concerns whether critical thinking is a general or a subject-specific skill. This debate is very important for this thesis, because it explores how the assessed construct should be perceived and justifies the decision why critical thinking is assessed as a general construct and a set of skills which can be applied in different contexts.

One of the most famous opponents of the idea that critical thinking is a general skill was McPeck. McPeck disagreed with the idea that critical thinking is a general construct and he claimed that critical thinking can only exist in a subject area. McPeck (1981) defined critical thinking as the 'appropriate use of reflective skepticism' (p.7), which means that critical thinking does not require scepticism in general. Critical thinkers should know when to ask questions and what the appropriate questions to be asked are. Critical thinking does not involve simply questioning or disagreeing with what is said. This happens only if it is necessary for a solution to be achieved or for the insight of a problem to be developed.

The key idea of McPeck about critical thinking, which probably distinguished him from Ennis and other scholars, was that 'critical thinking always manifests itself in connection with some identifiable activity or subject area and never in isolation' (McPeck, 1981, p.5). According to McPeck, somebody might be critical about X and not be critical about Y. McPeck stated that the study of logic (formal and informal) is not adequate for somebody to think critically. This is probably an argument that could also contradict the ideas of Lipman who argued in favour of the teaching of logic. Lipman (1987) wrote that the opinion of McPeck who perceived only a discipline-specific thinking has merit, but it is 'needlessly narrow' (p. 11).

McPeck identified two problems in the position of Ennis. The first issue is the contradiction between discussing for general critical thinking but using subject-specific dimensions. According to McPeck (1981), Ennis mentioned three dimensions of critical thinking: logical, criterial and pragmatic. He explained that even though Ennis

presented critical thinking as a general skill, only the first dimension incorporates logic, while the other two require specific knowledge and a particular subject area. The second issue is that, when critical thinking is perceived as a subject-independent construct, the statements discussed are always obvious and too generic to be useful. Specifically, McPeck (1981, p.52) mentioned that they 'typically degenerate into collections of near-tautologies or the most obvious kind of vacuous advice (for example, 'Select data that support your conclusion'; 'Do not contradict yourself')'.

It is important to clarify my own stance in this thesis. This thesis accepts that critical thinking is a general construct for various reasons. First of all, convincing arguments of why critical thinking can be a general construct has been presented. There are inter-disciplinary questions and the people cannot be experts in all of the subjects on which they are asked to take decisions. Therefore, students, who will be future citizens in a democratic society, will be required to decide on different topics and it would be impossible to always ask help from experts. Secondly, critical thinking is a general skill and according to Lipman (2003, p.44) despite the validity of some of McPeck's arguments, the existence of logic and philosophy as an independent discipline can prove that thinking can be perceived independently of disciplines.

This thesis argues in favour of critical thinking as a general construct which is also the stance that Lipman adopted (Lipman, 1987; 2003). This stance is also in line with the P4C programme targets. To be more specific, P4C dialogue does not focus on topics from particular domains or on perspectives which derive by specific disciplines. The questions are general, and the students can think about them without being subject experts. The programme itself adopts a general approach to critical thinking. For consistency, this is how it was assessed by this thesis.

Furthermore, Siegel introduced the necessity of two different types of principles for reasoning: 'subject-specific' and 'subject-neutral' (Siegel, 1988, p.34). This suggests that any scholar belonging to the Informal Logic movement or anyone who supported thinking as a subject-specific skill was partially correct.

This thesis is aligned to these ideas and to the ideas of Fisher and Scriven (1997) about the relationship between critical thinking and knowledge. Somebody might be an expert in a domain and this does not imply that they are also critical thinkers. On the contrary, a critical thinker might be able to pick controversies and errors, which might have been missed by the experts in that discipline. What is more, aligned to the positions of the same authors (Fisher & Scriven, 1997), it is accepted

that critical thinking can be applied in disciplines and this is compatible with the ideas of McPeck. However, it is also accepted that critical thinking can be accepted as a general skill, which has value and, hence, it is worth being taught and assessed even as a stand-alone subject. This general thinking skill can be applied with the common knowledge that everyone has as a citizen or on knowledge which is not necessarily linked to the official curriculum and disciplines taught in schools.

At this point, I judge necessary to add the two main arguments presented by Norris and Ennis (1989) in favour of evaluating critical thinking with general knowledge context. First of all, if critical thinking is evaluated in a general knowledge context, some students are not penalised if they do not hold knowledge in a particular subject. Secondly, if in reality general application of critical thinking on different contexts is desirable, then this is how it should be assessed. Therefore, I can argue that this type of assessment reduces the construct irrelevance. Furthermore, it provides an authentic assessment, which resembles the way critical thinking will be applied in real-life situations.

Paul (1985) argued that if McPeck accepts critical thinking as subject-specific because it is always 'thinking about X', then he should also reject the existence of the general ability of writing or reading. Even though there is writing about X or reading about X, it is possible for the students to learn and write or read in general. This is a strong argument and therefore I judge that it effectively suggests that critical thinking can be a general skill similar to reading and writing.

Finally, this thesis accepts the ideas of the critical thinking movement instead of critical pedagogy. It has probably become apparent that this thesis is related to the critical thinking movement and not the critical pedagogy. This movement has been criticized about the 'neutral' way that it perceived CT. Davies (2015) claimed that teaching critical thinking should not be considered neutral when social conditions are concerned and therefore it should include more than developing a set of abilities. Even though the sources related to critical thinking movement do not usually discuss critical pedagogy, it is important to explain the difference between the two and acknowledge this difference. Whilst the movement perceived critical thinking skills as neutral, critical pedagogy emphasised on the use of critical thinking as an important educational goal, which could help the students to change the status quo (Burbules & Berk, 1999).

I accept the critical thinking movement tradition at this stage, because I find this in line with the P4C tradition. P4C does not aim to guide the students to reach specific answers. On the contrary, it provides them with the tools to philosophise even if the dialogue remains open and there is no definite answer at the end of the session. This is in line with the critical thinking movement, which does not promote any indoctrination of the students in a particular ideology but suggests that the students should think for themselves and in a community. Finally, one of my main arguments when I introduced the significance of the study was the preparation of the students for the needs of the society and the economy. Critical pedagogy protects the student from the idea 'of being trained for the economic needs of large corporations' (Davies, 2015, p. 72).

Reed-Sandoval and Sykes (2017) argued that P4C should take seriously somebody's stance on an existing economic, political, cultural and social context. If this is accepted, then P4C cannot be neutral. For example, by not discussing racism or by discussing it in a 'neutral' way, it implies acceptance of whiteness and the status quo (Chetty, 2014). Nevertheless, I do not think that P4C has yet reached the type of sessions that critical pedagogy would have expected. P4C schools do always choose to reveal the possible oppressions of particular groups and leave the students to draw their own conclusions at the end of the dialogue.

Similarly, the critical thinking skills and assessments accepted by this thesis refer to a set of skills by accepting neutrality of these skills and without negotiating particular social stances. Although I do not extensively discuss critical pedagogy in this thesis, I do not reject it. However, I believe that a critical thinking assessment as perceived by the critical thinking movement is more appropriate when the effectiveness of P4C of students is examined and for the particular age group compared to an assessment of critical pedagogy.

3.2. Critical Thinking: Working Definition

In the previous section, it has been argued why critical thinking can be accepted as a general construct. In this part of the chapter, the working definition of critical thinking as a general construct will be discussed. The definition of critical thinking determines the way it is assessed (Butler, 2015). The common definitions of critical thinking included abilities and dispositions. In the section of definitions, the weaknesses of

assessing dispositions were presented. Therefore, dispositions were not included in the working definition for this thesis.

This thesis combined different definitions to establish its working definition. The definition was mostly based on Ennis (2015) because his focus on assessment resulted in a better operationalisation of the construct in his writings and on a Delphi report.

The Delphi report is an official attempt made to synthesise the ideas of the aforementioned and more scholars. Ennis, Lipman, Paul, Norris and other critical thinking experts were invited to a panel of experts by the American Philosophical Association (1990) in order to define the nature of critical thinking. This collective effort resulted in the production of the *Delphi report* with Peter Facione being the principal investigator. All the experts consent in one definition of critical thinking

> We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. CT is essential as a tool of inquiry. As such, CT is a liberating force in education and a powerful resource in one's personal and civic life. While not synonymous with good thinking, CT is a pervasive and self-rectifying human phenomenon. (p.3)

The definition also consists of dispositions for the critical thinker. The principal investigator Facione created a critical thinking test based on the dispositions that the report suggested. The test is named California Critical Thinking Disposition Inventory and it measures seven basic critical thinking dispositions: inquisitiveness, open-mindedness, systematicity, analycity, truth-seeking, critical thinking self-confidence and maturity (Facione et al., 1995).

However, in my measurement tool, I focused on skills and operationalised critical thinking as inference, evaluation of argument and particularly the examination of the credibility of sources, reasoning (and specifically deduction), assumption identification and problem-solving. I already demonstrated that reasoning and problem-solving are aspects included in prevailing definitions of critical thinking.

To be more precise, inference refers to the process of drawing a conclusion from certain observed or supposed facts (Watson & Glaser, 2002). Even though it is not apparent from the definition, Fisher and Scriven (1997, p.44) included inferences

as a part of critical thinking. According to the *Delphi report,* inference is a sub-category of critical thinking, and it includes three different sub-categories: querying evidence, conjecturing alternatives and drawing a conclusion (American Philosophical Association, 1990). The person should eventually apply reasoning to reach a conclusion which is most strongly warranted.

The Delphi report (American Philosophical Association, 1990) for critical thinking discusses the evaluation of an argument as one of the prime six aspects of critical thinking. This evaluation might incorporate judgement on the credibility of a source. The credibility of sources assessment involves the presentation of the statements from different people and the student should judge whether the advice is credible. Fisher (2010) divided the process of judging the credibility of sources in five elements: the source whose credibility is judged, the context, the justification the source offers, the nature of the claim and the association with other sources. Particularly, when Fisher (2010) discussed the first element, which is the person or the source whose credibility is judged, he mentioned a few secondary questions: whether they have the relevant expertise, whether they have the ability to observe accurately, whether their reputation suggests they are reliable and whether in the particular context the source might be biased.

The working definition of critical thinking includes reasoning and deduction. Many scholars included reasoning in critical thinking (Ennis, 2015; Paul, 1993; Siegel, 1988) and the Delphi report is also compatible with this (American Philosophical Association, 1990). Even though it might seem self-evident that deduction is a part of critical thinking, there are researchers who did not support this stance and distinguish deduction from critical thinking. For example, Newton (2014; 2015) talked about productive thinking and according to his categorization deductive thinking is another type of thinking which is a sub-category of productive thinking – amongst creativity and critical thinking. I disagree with this idea. I believe that both deductive thinking and induction should be included in the critical thinking. *Harry Stottlemeier's Discovery* is the first novel Lipman wrote for P4C sessions and he introduced the students to Aristotelian logic (Splitter, 1992) and therefore to deductive reasoning. I consider deduction the most problematic way of thinking because of the criticisms it has received. Specifically, Evans (2005, p.169) asserted that false premises can draw a valid conclusion. In the chapter that I discuss the construction of the measurement

tools, I will include a discussion to articulate the consideration of this limitation of deductive thinking. I will explain how this particular type of thinking was evaluated.

Assumption identification is the fourth aspect of critical thinking used for its operationalisation. According to Brookfield (2012), the basic process of critical thinking involves assumption identification in thoughts and actions, evaluation of the validity of these assumptions, looking ideas and actions by multiple perspectives and taking informed decisions. Assumption identification and evaluation play a crucial role in the critical ability of a person.

Finally, problem-solving was included as an aspect in the critical thinking assessment. Previously, it was stated that Sternberg (1986) and Halpern (1998) emphasised this aspect. One of the reasons that I argue that critical thinking is vital to be enhanced by the school is because of the applicability of this skill in the everyday life of students and their future life as adults and citizens.

To conclude, this definition chose particular aspects of critical thinking definitions and can be considered to be closely related to the ideas of Ennis about critical abilities (Ennis, 2015). Nevertheless, not all the abilities of Ennis were chosen to be evaluated. This is not because they are not considered a part of critical thinking, but because it is believed that critical thinking can be sufficiently operationalised for the purposes of this thesis without including all the skills that Ennis included in his updated definitions. This recommendation is also made by Fisher and Scriven (1997, p.85) when they discussed the critical thinking definition suggested by Ennis: ‘One must, however, be careful not to include everything on it as relevant to a test of critical thinking without careful thought: it casts the net too wide by a mile’.

3.3. Creativity: Definitions

The second construct to be evaluated was creativity. The work of Wallas is the starting point for the creativity research. Even though his work seems to discuss problem-solving, Wallas is in fact focused on the ‘art of thought’.

According to Wallas (1926), there are four stages of thought. First, there is the stage of ‘preparation’. At this stage, a problem is investigated in all directions. This is a conscious process. As he discussed, the mind gives a clear answer only where there is a clearly set question for which evidence can be sought. The second phase is called ‘incubation’ and during that phase the thinker is not actively engaged in this problem.

During that phase, the thinker does not voluntarily think of the problem, but there are a series of involuntary mental events taking place. The thinker can spend this phase either by working on a different problem or by relaxing. There are two things that Wallas discussed and should be included in this phase. First, during the phase of incubation, physical exercise might take place. Secondly, the danger of this phase is that the thinkers keep being involved in reading and therefore they keep thinking of the question.

The third phase is called 'illumination' when a new idea occurs unexpectedly to the thinker. Finally, the fourth phase includes the 'verification'. This involves conscious thinking. During this phase, the new idea is tested and its consequences are considered.

This process described by Wallas is a clear description of a problem-solving activity. In this thesis, problem-solving was included in critical thinking instead of creative thinking. However, the generation of ideas which are related to a particular problem can be considered creative problem-solving. I consider the model of Wallas very important for the examination of creativity studies because it involves the generation of new ideas. Nevertheless, I recognise a limitation. Creativity was mostly associated with intelligence and the studying of creativity focused on the research with genius children and adults. Therefore, I believe that Wallas discussed a process which is more complicated and at a higher level than daily creativity and requires some time to be expressed. If every simple creative action in daily life required an incubation period, it would have been impractical and time-consuming for people's lives.

3.3.1. Definitions

There are different definitions of creativity and there is no consensus. The *Cambridge Handbook of Thinking and Reasoning* has a chapter specifically on creativity. Sternberg, Lubart, Kaufman and Pretz (2005) summarised all the different types of literature related to creativity which are currently available. First, there is the mystical approach to the study of creativity, according to which a divine or Muse might inspire the creator. Sternberg and Lubart (1999) offered examples of a pragmatic approach and mentioned examples of techniques which could help the students to become more creative, such as the technique of brainstorming and the technique of removing the perception that there is only one right answer. Sternberg et al. (2005) mentioned other approaches. Specifically, they reported the psychodynamic, psychometric, cognitive,

social, evolutionary, confluence and alternate approach. In addition, Plucker and Renzulli (1999) referred to five categories of creative studies: psychometric, experimental, biographical, biometric and historiometric.

The categorisation offered by Sternberg et al. (2005) is more extensive. Its main benefit is the fact that the authors also offer the taxonomy of creative contributions. They discussed different types of contributions, such as replication of a study or redirection of a field. The question was not restricted to whether a process can be judged creative, but what type of creativity is demonstrated by an individual. Plucker and Renzulli (1999) categorised the literature in a less detailed way but I think their categorisation is more straightforward.

This thesis can be considered as what Plucker and Renzulli called experimental approaches with creativity since this thesis will mostly involves measurement tools of creativity, a comparison and an intervention group as the psychometric approach would have suggested. Psychometric approaches are concerned with the person, the product, the process and the environment. It is apparent that different factors, assessments and definitions are considered for each of these approaches.

3.3.1.1. Person

When there is a reference to the creative person, there is usually an attempt to search for related personality characteristics. Davis (1999) summarised the personality traits of creative people. Creative people are meant to be aware of their creativity. They are original, independent, willing to take risks, energetic characterised by thrill-seeking, thorough and curious. They have a sense of humour and capacity for fantasy. They are attracted to ambiguity and complexity, artistic, open-minded, perceptive, emotional, ethical and they need time alone. The one I believe needs clarification is the element thorough, because it demonstrates that creative people are actually self-disciplined, organised and perfectionists. Creative people do not simply wait for inspiration as the mystical approach to creativity would suggest. One of the personality traits of creative people is being hard working. Furthermore, Davis (1999) reported the negative traits of some creative people, such as their childish, neurotic or even slightly sociopathic behaviour.

Amabile's work also focuses on motivation as a characteristic of creative people. Amabile (1995) conducted an experiment trying to identify the impact of motivation on creativity. She found that the poems produced by adults after having

responded to an extrinsic orientation were less creative compared to those produced by adults who had just responded to an intrinsic orientation questionnaire.

According to Piirto (2010), some core attitudes for the creative process, are tolerance for ambiguity, self-discipline, risk-taking. Furthermore, she mentioned openness, which according to Piirto (2010), is the ability of creative people to be curious and pay attention to small things. These elements have already been discussed by Davis (1999) and in what follows it will become apparent that they are also aligned with Torrance's views.

In Piirto's pyramid of talent development (Piirto & Ford, 2000), there are many personality traits which are in alignment with those discussed for creative people, such as intuition, openness, passion for work, perceptiveness, perfectionism, risk-taking, tolerance for ambiguity. There are also some additional traits, such as androgyny, perfectionism and resilience.

To conclude, I did not identify extremely contradictory references when the characteristics of creative people are presented. Some sources are complementary to others, but the personality traits presented are usually aligned. Considering what was previously discussed with the dispositions for critical thinking and the distinction between the critical thinking and the critical thinker, it can be argued that somebody might have creative personality traits, but they might not generate creative outcomes.

3.3.1.2. Process

Guilford (1950) discussed creativity and he focused on the creative abilities. He clearly stated the abilities are those that can determine whether the person will be able to display creative behaviour. He presented a list of creative abilities. As he explained, these refer to the creative abilities needed for scientists, inventors or people who are focused on technology and do not necessarily apply in other domains. These are summarised as:

- Sensitivity to problems, which involves skills such as asking questions
- Fluency
- Flexibility
- Having novel ideas. Regarding this, Guilford (1950) clarified that in order for these novel ideas to be considered creative, they should be acceptable.
- Synthesising ability

- Analysing ability
- Reorganisation of organised wholes
- Resistance to confusion and able of handling complex mental structures
- Evaluation

Guildford (1950) suggested this definition as mostly appropriate for particular domains, but I suggest that this definition can be appropriate for creativity as a domain-independent skill. Even though evaluation is adopted as an element of critical thinking by this thesis, all the other elements are accepted for creativity as a general construct and hence most of these elements will be used in the working definition of this thesis. The most important element of this definition is that the generation of creative responses does not have to come from anything. Synthesising and analysing existing ideas can result in a product which can be considered novel and creative.

Also, Guilford (1956; 1967) included divergent production as one of the six operations in his Structure of Intellect. Particularly, he suggested that divergent production can refer to units, classes, relations, systems, transformation and implications and it might be on a figural, symbolic semantic or behavioural level (Guilford, 1967). In an earlier version of the system of intellect, Guilford (1956) referred to the particular type of content in divergent production. He explained that the production might refer to production in words, ideas, expressions, shifts, novel responses and details. As a result, he used the term 'flexibility' for the shifts, elaboration for the details, the term 'originality' for the novel responses and 'fluency' for words, ideas and expressions.

In what follows, it is necessary to refer to the definition of Torrance. It can be supported that Torrance continued the tradition from Guilford and expanded on it. Even though Torrance discussed various definitions for creativity, he summarised his research definitions in what follows (Torrance, 1988, p. 47):

> 'I tried to describe creative thinking as the process of sensing difficulties, problems, gaps in information, missing elements, something askew; making guesses and formulating hypotheses about these deficiencies; evaluating and testing these guesses and hypotheses; possibly revising and retesting them; and finally communicating the results.'

There are many interesting aspects included in Torrance's definition. First of all, he clearly discusses creativity as a process. Secondly, he discusses both problem-finding and problem-solving as parts of the creative process. It also includes an audience in the definition. Moreover, the definition of Torrance includes elements which were discussed in the definitions of critical thinking earlier in this chapter, such as hypothesis testing. This should not be surprising, because of the idea of critico-creative thinking (Fisher, 2010) which was also introduced earlier in this chapter. Even though this thesis discusses these two concepts separately, both concepts can be considered examples of thinking skills and related - at least to some extent.

Lipman (1987, p.10) claimed that critical thinking and creativity are 'compatible and even overlapping'. Norris and Ennis (1989) described the relationship between critical, creative and good thinking. They said that critical thinking is a part of good thinking and it can be separated into evaluative and non-evaluative thinking. Likewise, creative thinking is a part of good thinking and includes reflective and non-reflective thinking. These two parts overlap when the thinking is reasonable, reflective, productive and non-evaluative. The first two elements are included in the definition of critical thinking in all cases. However, when the thinking is reasonable and reflective, but also evaluative and non-productive, then it is also an area of critical thinking and it does not overlap with creative thinking. According to Norris and Ennis, creative thinking is always non-evaluative, productive and reasonable.

Burbules and Berk (1999) perceived creativity as an alternate version to criticality. They emphasised that criticality does not only include finding a meaning, but also creating a meaning and to think in a different way. They focused on elements that they have been emphasised as elements of creativity for a long time: openness and imagination. Although these authors refer to criticality and the social character of thinking, which is not necessarily how critical thinking is perceived and defined by this thesis, it can still become obvious that they recognize the close link between criticality and creativity.

Fisher and Scriven (1997) also attempted to set a relationship between critical thinking and creativity. They referred to the creativity that exists in critical thinking, which is different from what is usually mentioned as creativity. They said that this type of creativity has different characteristics. It operates with language and not art (e.g. dance or painting) or mechanical invention. It requires novel ideas. These ideas should be novel for the specific context. Therefore, novel is not referred to the novelty that a

Nobeler would require. They named this type of creativity 'functional creativity'. This is the type of creativity that this thesis focuses on. It is not the creativity that should be related to arts, science or a specific discipline. It is the creativity which is linked to (critical) thinking.

The creative process requires some intentionality. Craft (2001) argued that fantasies and dreams are not creative here because there is no conscious intention to create them. I only partially agree with Craft. I agree that fantasies and dreams are not typical creative products. However, this is not because of lack of intentionality. Fantasies and dreams still have a creator and this is the person who imagined them. Furthermore, even in Wallas (1926) illumination, the mental activities were not intentional. I argue that fantasies and dreams refer to a creative process but they are not creative products. In other words, fantasies and dreams are characterized by creativity. However, they are not creative products because they can only be experienced by a creator and they do not have an audience. As soon as these are used in order to be the basis of a different product, such as an oral story, which is communicated to an audience, they can lead to a creative product.

3.3.1.3. Product

Runco and Jaeger (2012) wrote what is called a standard definition of creativity. According to them, creativity requires originality and effectiveness. This means that original products should be uncommon, and they should have a value. Corazza (2016) accepted these two elements but suggested that this definition perceives creativity as static. Creativity should be perceived as a dynamic process. Therefore, creativity requires potential originality and effectiveness. A creative thinking process sometimes might not produce a creative product or reach a specific conclusion. In that sense, a creative agent is the one who pursues and not necessarily achieves creative goals.

By the examination of what Corazza (2016) argued in relation to creativity, it becomes apparent that there is no concentration on the product itself, since in some of the cases there is no production of a conclusion or a final product. I think the dynamic definition is useful in a theoretical framework. However, it might be problematic used for educational assessments. The potentiality in the definition of creativity turns the focus to less visible elements of the creative process. It is difficult to distinguish who is more or less creative in a dynamic definition of creativity. From a pedagogical point of

view, it is useful to see potential in students. It is difficult, however, to make concrete judgments about this.

The standard definition of creativity was also criticised on other aspects. Weisberg (2015) questioned the inclusion of value in the definition of creativity. The main argument is that the judgment about the value of a creative product is extremely subjective. As a result, he suggested the definition that a product is creative if it is novel and produced intentionally. This working definition adopted by this thesis is in agreement with the recommendation of Weisberg (2015) for the exclusion of value, because this might lead to subjective assessments which may be highly culturally and time-dependent.

Similarly, James and Taylor (2012) criticised the element of usefulness required in order for a product to be judged as creative. They explained that it has to be questioned whether this aspect refers to the usefulness for the agent, but not for the other people. For instance, a robbery might be useful for the robber, but not for the people who are being robbed. Therefore, James and Taylor (2012) argued that negative creativity should be distinguished from creativity which has unintended negative consequences. Their argument, however, still suggests that the usefulness or the value of a product is not an objective indicator. Something which is valuable for someone for the time being might not be for others at the same or different time.

3.3.1.4. Environment

The environment is also an element widely discussed in creativity literature. Many researchers argued that the environment can foster creativity. The acceptance of the belief that the environment can support creativity is important for this thesis. This means that creativity can be developed and interventions in the life of an individual can have an impact on creativity. Therefore, it can be questioned whether P4C can be one of these interventions.

It is useful to present an example of a definition in which the environment plays a crucial role. Plucker, Beghetto and Dow (2004, p.90) defined creativity as '[...] the interaction among aptitude, process, and environment by which an individual or group produces a perceptible product that is both novel and useful as defined within a social context'. According to this definition, environment is needed both as a factor to influence creativity and to set the context in which creativity is defined. As it will be presented later in the grading process of the creativity, even though cultural knowledge

was not rewarded in the responses of the students, there is an acceptance that cultural knowledge and context was used in order to grade the responses and categorise them.

Also, the environment added a different perspective to this thesis, since the school can be considered an environmental factor. Piirto and Ford (2000) presented the pyramid of talent development and included five suns above the pyramid, which could potentially affect the development of talent. These suns stand for the home and the family, the school, the community and culture, the gender and chance. Even though the presentation of gender might seem bizarre in a model of talent, this is not a sexist stance of the authors. The authors simply accepted that even though both boys and girls are born with the same talent, gender might play a role in the society. Thus, it will impact on how people develop or how much they are rewarded for their talents. Similarly, the community plays a crucial role in whether the talents of a person get recognised. This thesis concentrates on the influence of the second sun and examines whether the school and particularly a school-based intervention can play a role in the development of pupils' creativity.

Amabile (2017) discusses the impact of the environment on creativity. This is because the environment can have an impact on motivation. Since the environment impacts on motivation, motivation can also have an impact on creativity. This can bring a different approach to the current research because a potential interpretation of any impact that P4C plays on creativity might be explained through motivation. For example, the positive or negative impact that P4C could have on creativity may result from a change in the motivation for learning of pupils instead of being a direct effect on their creativity.

To summarise, in this thesis there is an experimental study to examine whether a school-intervention as an environmental factor can have an impact on the creativity of the students. Furthermore, the environment and the social context were also used to some extent as references for the grading of the creativity activities in the assessments.

3.3.2. Is creativity value-neutral?

There are examples of behaviour which can be considered creative, but not ethical. This might set the question whether there is a dark side of creativity. It has already been mentioned that the negative creativity should be distinguished from the creative actions which have unintended negative consequences (James & Taylor, 2012).

Runco (2012) argued that creativity has no dark side as a process. The intentions and decisions which direct the process can be malevolent, but they are not synonymous with the creative process. The product might be malevolent. He explained that creativity is always deviant, and this is the reason why creativity might be perceived as malevolent.

Sternberg (2012) accepted that there is a dark side of creativity, but there is a way to reduce this. In order to distinguish different types of creative actions, such as those which aimed to increase the common good, the actions of Hitler and actions of self-interest, Sternberg (2012) argued that creative actions should be characterised by wisdom. This means that creativity should achieve the common good by balancing intrapersonal, interpersonal and extrapersonal interests. When creativity is perceived in this way, it is characterised by universal values accepted by ethical systems around the world. Sternberg (2012) recommended a school intervention to develop this wisdom. Particularly, he said that if the students develop dialogue related to literature and philosophy, they can develop their wisdom. Even though he did not explicitly refer to P4C, this intervention appears similar to P4C.

This thesis accepts that creative products can be malevolent or not, but it does not aim to evaluate any of creative products ethically. This thesis is focused on creative thinking process. The responses of the students which might have negative consequences are also considered creative. For example, using a brick as a weapon is still judged as a creative response. There is no evaluation of values, intentions or consequences of the creative process.

3.3.3. Is creativity a domain-specific skill?

The question of whether creativity is domain-independent or domain-specific was also set, as in the case of critical thinking. Even though arguments were developed for both sides, Craft (2005) stated that in the UK creativity is accepted as a 'generalised phenomenon' (p.15) continuing Guildford's tradition. The main difference, however, is that even though the initial tradition of creativity focused on genius, nowadays the focus is on ordinary creativity. This is what Craft (2001) called 'little c in creativity' or what Fisher and Scriven (1997) called 'functional creativity'. This is the stance adopted by this thesis.

This thesis does not examine the creativity in relation to a specific discipline. However, this does not mean that creativity develops in a vacuum. Craft (2001)

specified that creativity is always in relation to something. This is not contradictory with the stance of this thesis, which discusses ordinary creativity which is independent of a domain. Craft (2001) argued that creativity needs a context to be developed and creativity develops in relation to something. Indeed creativity does need a context. I argue that the need for context is actually dual. Creativity needs context in order to be expressed and it needs a social context to be judged. Creativity is expressed within a specific context and its product can only be judged as creative within a context.

However, I argue that even though creativity requires a context to be expressed, it is not restricted to this context. Ordinary creativity can be transferable. If it is accepted that creative individuals have different personality characteristics than non-creative, then these personality characteristics would enable creative behaviour in everyday life in various contexts.

Additionally, I argue that if creativity is domain-specific, then it is assumed that knowledge is always required in order for somebody to be creative. Nevertheless, there is no clear linear relationship between knowledge and creativity. It has been argued that knowledge experts can have a fixed way of thinking which impedes them from being creative (Sternberg & Lubart, 1999). As a result, even though some knowledge might be required, people's knowledge in a domain and their creativity are not proportionally developed.

3.4. Creativity: Working Definition

As it has already been discussed, the current thesis accepts creativity as a subject-independent construct. The definition I used in order to operationalise creative thinking remains close to the psychometric tradition and adopts most of the elements from there.

Creative thinking is defined as the generation of ideas which are innovative and imaginative. Creative thinking requires elaboration of ideas, ability to abstract their essence and openness to the vagueness that is required during the creative process. Even though the definition adopted by this thesis focuses on creative thinking as a process, in fact these creativity elements are evaluated via the outcomes and the responses that the students provided in the assessments.

To specify these elements, initially there is a reference to divergent thinking. Guilford (1967, p.233) defined divergent production as 'generation of information

from given information where the emphasis is upon variety and quantity of output form the same source; likely to involve transfer'. Particularly, the first part of the working definition of this thesis is concerned with the generation of ideas involving both fluency and flexibility. According to Guilford (1967, p.138), fluency is an ability which refers to the 'flow of ideas' and flexibility refers to the 'readiness to change direction or to modify information'.

Hence, the first part of the definition is mostly based on the tradition of Guildford who focused on this divergent thinking. Even though this part of the definition focuses only on quantity, there is also the element of qualitative evaluation in the definition adopted by this thesis. The students do not only have to generate ideas, but they should also have to generate ideas with some novelty. Standard definition of creativity included novelty. Torrance, Ball and Safter (2008) named the element of infrequency of answers 'originality'. This thesis accepts that the terms novelty, innovation and originality can be used interchangeably.

According to Craft (2001) there is no creativity without innovation or novelty. However, she sets a very important restriction about this element. Particularly, Craft (2001, p. 56) used the phrase 'doing it differently'. This is the phrase that is accepted as novelty from this thesis. There is a comparison amongst the students in order to identify who is suggesting different ideas. In that sense, the novelty element evaluated by this thesis is context and sample dependent. Even though this might sound like a restriction, this is not the case. When innovation or novelty is discussed the agent should be considered. For example, something might be novel for a child but not for a field of experts.

Despite evaluating the 'functional creativity', there is no evaluation of its functions in a narrow sense. It is not accepted that a creative idea has to be useful. A creative idea in everyday life does not have necessarily to be useful in a practical sense. For example, a humorous response might satisfy an emotional or psychological need without being useful for daily life. Alternatively, it could be argued that entertainment is a function of 'functional creativity'. Moreover, in agreement with Weisberg (2015), it would have been very difficult to achieve an objective measurement of the value of the response, and therefore the value and the effectiveness of the answers of the students were not assessed.

Elaboration, abstractness and resistance to premature closure are three elements also assessed by the Torrance Tests of Creative Thinking (Torrance, Ball & Safter,

2008). Elaboration refers to the detail added by the student and Torrance Tests of Creative Thinking linked that with imagination. Abstractness refers to the process of thinking and the abilities to analyse and synthesise. Abstractness, however, refers to the best form of these two abilities and it is what enables the person to catch the essence of the ideas and to separate the significant information from the trivial. Resistance to premature closure refers to the ability to remain open and not rush into conclusions. Davis (1999) who discussed the personality traits of creative people also suggested that creative people have similar personality characteristics. They are open-minded and attracted to complexity and ambiguity. Therefore, when these three elements are concerned, the behaviour that the participants of this study demonstrate during the creativity activities can also be considered an indirect assessment of these personality traits.

Finally, the working definition of creativity includes an element of imagination. According to Craft (2001) imagination enables somebody to see more than what is evident in the first place. The approach that Davis (1999) presented imagination is very interesting. He mentioned that this term is a complex one and he used the term 'visualisation' and 'synesthesia' to explain this. Visualisation is the ability that somebody has to see something in their head, fantasise and manipulate images and ideas. The term synesthesia suggests that this does not apply only to images, but also to sounds and other senses. For example, Mozart imagined compositions.

Imagination is the only element in the definition not evaluated directly in the assessment of this thesis. This is due to the fact that I consider imagination an internal creative process. Imagination is important for creative thinking and it is assessed indirectly by examining the creative products occurred. The creativity activities of this thesis require imagination in order for the students to generate ideas.

3.5. Can critical thinking and creativity skills ever be improved?

This thesis examines whether P4C can develop creativity and critical thinking. However, whether these skills are malleable by any intervention can be questioned. There should be evidence that there are other interventions which can improve these skills before examining whether P4C can change them.

Critical thinking skills can be considered malleable because there is currently evidence suggesting interventions which can improve students' critical thinking skills at all education levels, including primary education (Abrami et al., 2008). Some

evidence focused on the development of these skills in college students (Kong et al., 2014; Niu, Behar-Horenstein & Garvan, 2013). However, as Lipman suggested, it is also important to focus on the development of these skills in students of younger ages. Furthermore, as Abrami et al. (2008) reported, studies who engage primary school students as participants are these which report bigger effect sizes compared to those with college students. This finding may suggest that these skills can be easier improved in younger ages. Studies have already examined the impact of different interventions on primary schools students. For example, a recent study with students in Hong Kong reported that effective group work had a positive impact on primary schools pupils' critical thinking (Fung, 2014).

Concerning the evidence about creativity development, there are some recently published studies which suggest a change in creativity performance after specific interventions. For example, a study in New Zealand examined the impact of project-based learning on creativity (Storer, 2018). Grade 4 students participated in this study. The intervention lasted only for six weeks and only 90 participants were involved in the study. The impact of the programme on fluency, originality, elaboration, abstractness of title and resistance to premature closure was measured. The definition of creativity used in that intervention matches to the working definition of this thesis. I used the reported means and standard deviation in the pre-test and post-test in order to calculate the effect sizes between the groups (which were not reported by the researchers) and I found that positive effect sizes were found in most of the areas (fluency = -0.1, originality and resistance to premature closure = 0.1, Abstractness of title = 0.06, whilst the effect size of elaboration was calculated as 0.5). Therefore, small or medium positive impact was found in some creativity areas, which suggests that these skills can be developed.

A different study also offered evidence that the Skills4Genius Programme can lead to creativity enhancement for the intervention group. Positive effect sizes were reported for elaboration, originality, premature closure and abstractness to titles (Santos et al., 2017). No clear finding was reported in relation to the impact on fluency. The intervention in this study lasted for five months. However, the sample was smaller than the previously mentioned study with 22 participants in the experimental group and 18 in the control group.

See and Kokotsaki (2016) also identified some studies which suggested that arts education can enhance students' creativity. The researchers expressed concerns

about specific weaknesses in the design of the studies. There is also evidence that interventions improve the creativity of adults. For example, a study with 53 female students showed that the engagement in game making led to increase in divergent thinking scores (Gallagher & Grimm, 2018). Another example comes from a recently published study in United Arab Emirates reported that a training on creativity, which was not discipline-specific, enhanced the creativity of the participants (Vally et al., 2019). However, this study did not have a comparison group, so its research design is weak to establish a causal relationship between the intervention implementation and the change in creativity scores.

To summarise, there is no consistency in the existing evidence to secure that these skills are malleable. Some of the evidence suggesting that these skills are malleable comes from short-term interventions with small sample. No rapid change would be expected in these skills and studies with a bigger sample are needed to establish that these skills are malleable. Considering potential publications bias, there might be a tendency for interventions to report some positive results. Therefore, it can be questionable to what extent the results of these studies are trustworthy and generalisable. However, since there is some available evidence which suggest that a difference in these skills could be expected after an implementation of an intervention, a similar result might occur after the implementation of the P4C intervention. This is an encouraging finding. Furthermore, it would be a very pessimistic approach for the education to accept that there are no interventions to improve students' thinking skills. Even if there was no known intervention improving these skills, educational researchers should investigate this possibility because of the importance of these skills in later life.

3.6. Chapter Summary

This chapter discussed several definitions of critical thinking and creativity in order to support the two working definitions used in this thesis. The reasons why these two concepts are accepted as domain-independent were explained. Presenting working definitions for this thesis was crucial because the operationalisation of these concepts enabled their evaluation. Before concluding this chapter, it is also crucial to reiterate that the definitions used by this thesis are not exhaustive and they do not cover all the necessary elements of these two concepts. Hence, these definitions are not presented as the ideal definitions of these concepts. The definitions adopted by this thesis prioritised

specific elements of these concepts which were judged significant and assessable. This prioritisation was important in order to design or use assessments which could fit the age and the concentration span of the students who participated in this research. Finally, this chapter showed that there is some evidence suggesting that students' critical and creative skills can be developed following interventions.

4. Methods: Systematic Literature Review (Research Question 1)

The first research question of the study examined the impact of the P4C on cognitive and non-cognitive domains based on the existing published evidence. This question focuses on searching for the domains and skills that the intervention seems to have a positive impact on and how big this impact is.

As Gorard, See and Siddiqui (2017) recommended evidence in education can be envisaged as a cycle. This cycle may start with an evidence synthesis to demonstrate which questions remain unanswered and then primary research may be conducted. This is how this first research question informed the following research questions of this thesis.

4.1. Research Design

The first research question of this thesis examines the existing evidence of the impact of the programme. The popularity of P4C across the world has led to the design of various research projects to investigate the effectiveness of the programme. Therefore, it could be argued that evidence should exist to support the programme impact.

Different research projects focused on P4C impact on cognitive and non-cognitive domains by using a range of research designs and assessment tools. In some cases, the teachers were asked to identify the claimed benefits of P4C for their students. For example, teachers in Liverpool were asked to report P4C outcomes (Meir & McCann, 2017). Action research conducted in New Zealand to investigate the extent that P4C can contribute to thinking development, critical thinking and questioning of the students over a seven-month intervention (Benade, 2011). Similarly, a study (Green & Cody, 2016) conducted in South Africa, with focus group interviews, asked final year university students who experienced P4C to evaluate its benefits. In a study conducted in Greece, P4C gains were assessed by analysing student discourse (Gasparatou & Kampeza, 2012).

These and many more studies reported positive impact of the programme on various domains. For instance, studies reported positive impact on academic, behavioural and social domains (Meir & McCann, 2017), student reasoning, collaboration and democratic principles (Green & Cody, 2016) and high-order literacy thinking and language skills for four poor readers (Jenkins & Lyle, 2010).

To sum up, there is research evidence suggesting the positive P4C impact on cognitive and non-cognitive domains. Nevertheless, not all evidence is of the same quality. Without underestimating the research design of the aforementioned projects, experimental design and the studies with a comparison group can be judged as an appropriate research design which could fit the causal question of the P4C impact on a skill or knowledge domain. An experiment is recommended to establish the relationship between an intervention and its impact (Cohen, Manion & Morrison, 2007). Furthermore, Lipman (1987) recommended experimental designs in projects in order to distinguish effective programmes from the ineffective.

For this reason, the first research question was not answered by collecting any evidence about the P4C intervention. Instead, a systematic literature review with inclusion criteria was conducted. This review focused on experimental studies, quasi-experimental studies and studies with a comparison group. Consequently, this study evaluated the P4C impact on cognitive and non-cognitive skills by focusing only on studies whose research design is judged suitable to justify claims between an intervention and its impact.

4.2. Inclusion criteria in the systematic literature review

The search of studies was conducted primarily by using Google Scholar supplemented by hand-searches, expertise and snowballing. After some literature was retrieved, more literature was pursued based on the bibliography found. The main inclusion criteria for the studies in this systematic literature review were the research design and the purpose of the study. For a study to be included in the literature review it should:

- have had an experimental design, quasi-experimental design or at least a research design with a comparison group
- have included the conduct of both pre-test and post-test
- have examined the impact of P4C on one or more skills
- have been published in English from 1982 to 2018. An earlier version of this review with studies published 1982- 2017 has been published (Ventista, 2018b).

4.3. Scale for the evaluation of the quality of the controlled trials

Two of the four criteria for the inclusion of the study were related to their research design. Even though the existence of a comparison group and the pre-test and post-test

assessment enables the investigation of the counterfactual cases and examining the pre-test equivalence of the two groups, the studies included in the systematic literature review did not provide equally trustworthy results.

The design varied amongst the studies. Therefore, there was a stage of evaluation of the trustworthiness of the studies. This scale was created based on the idea of judging the trustworthiness of the studies suggested by Gorard (2015b) and Gorard, See and Siddiqui (2017) and finally by the scale used in Siddiqui and Ventista (2018). According to their estimation of the trustworthiness of the study, the researchers considered the strength of the research design in relation to the research question, the scale of the study, dropout, data quality and other threats to validity.

I created a system based on these recommendations to evaluate controlled trials which examine causal research questions. However, I did not follow the scale suggested by the authors completely because I focused on the consistency of rating and replicability of my findings. The scale suggested by the authors is intuitive, which means that a lot of times the person who evaluates the study has to take serious decisions without clear thresholds. For example, about the scale of the study the authors recommend large number of cases for the highest rating, medium number of cases for the next category. It becomes obvious that this is not a clear threshold, and this is a relatively vague judgement. Without the assistance of a rubric, different raters might consider the same study as having high or medium number of cases.

According to the system I developed (Table 4.1.) there are three different areas to be considered. The indicators of the quality of the studies are symbolised with stars. Each of the three areas offers to the study particular numbers of stars. The maximum number of stars that a study can get is 5, while the minimum is zero. Each category can offer a different number of starts in the final grading.

This grading system refers to the specifications of the research design and reporting of the studies in this systematic literature review. It grades only three areas suggested by Gorard, See and Siddiqui (2017). These authors evaluated the research design based on the research question. However, the system I developed refers to only controlled trials aiming to respond to causal research questions. First of all, this system rates the research design of the any controlled trial based on the way that the participants were assigned to a comparison or intervention group. If that was random then the study gets two stars for the overall score. If there was matching based on specific criteria, then the study gets one star because it is recognised that the matching

was based on known criteria and this assignment to group might have not been effective if the criteria were different. Finally, a study receives zero stars in case there was only a comparator group.

When the number of the participating units in a study is very small, then the randomisation cannot be considered trustworthy (Gorard, 2013, p.128). For this reason, I did not give two stars to studies with small sample even if they claimed random allocation of participants within the groups.

The second indicator of quality is the sample size of the smaller group. The idea of evaluating the N of the smaller in number comparison group is based on the scale of Gorard, See and Siddiqui (2017) who evaluate the sample size based on comparison group. However, as it has already been mentioned, the same authors did not set a clear threshold for the sample size to distinguish high, medium and low effect sizes. I decided on the number 100. This is arbitrarily, but I chose to use if for consistency reasons and in order to make my results replicable. However, I recognise that a study which might have 99 cases in the smaller comparison group does not significantly differ from the one which has 100. However, a threshold had to be set somewhere and if I enable a study with 99 cases to be considered in the other category, the same argument could apply for a study with 98 cases and so on.

Finally, the third indicator was the attrition of the study from the pre-test to the post-test. A study which does not report dropout should be graded with zero stars, because it is untrustworthy. A study which reports attrition, which is higher than 15% of the overall sample, introduces serious concerns about the results. Hence, this study is rated with one star. The study which reports attrition, but it is smaller than 15% of the overall sample, it can be graded with two stars. As it applied in the threshold for the sample size, 15% is an arbitrary threshold that was applied for consistency in the grading and enabling the replicability of the study. Furthermore, one main weakness of adopting this approach is the fact that the attrition of both groups is considered as a unity. In some cases, participants drop out form both groups whilst in others participants drop out only from the one.

Table 4.1. Trustworthiness Indicators for Evaluation of the Research Design of the Studies evaluating the impact of the programme.

Trustworthiness indicators	★★	★	0	Total Marks per indicator
Research Design	Randomised Controlled Trial with big sample (at least 100 participants in each group)	Matched Comparison group or randomisation within groups (with smaller sample)	Comparator group	0-2
Sample Size (of the smallest group in the study)	Not applicable	N ≥ 100	100 > N	0-1
Attrition (from pre-test to the post-test)	Reported attrition which is ≤ 15 % of the overall sample	Reported attrition which is > 15% of the overall sample	Not reported attrition	0-2
Total Stars for the Study:				0-5

This evaluation system is not exhaustive, and these are not the only indicators to be used for the evaluation of the studies. There are other criteria which can reduce the trustworthiness of the findings, such as the measurement tools used. A measurement tool which is focused on the exact skills targeted by the school-based intervention is more likely to demonstrate bigger impact for the intervention group. Similarly, the pre-test equivalence can play a significant role in the results that occur, and it was not examined by the grading system. These were excluded from the general scale because they would make it excessive and overcomplicated. These additional characteristics are included in the discussion and the judgment of the studies individually.

This system of quality for the study does not demonstrate anything about the impact that the study found about P4C. A high-quality study can find any type of impact (positive, negative or no impact). The impact of the programme should be examined separately from the quality of the study. In this systematic literature review,

the impact of the programme was examined after the inclusion and the evaluation of the studies.

4.4. Impact

The purpose of this systematic literature review was to reply to the research question on the impact of P4C on cognitive and non-cognitive domains. For this question to be replied, firstly the knowledge or skills where P4C has an impact were considered. Secondly, the impact that these studies report on these domains or skills were examined. The latter involved the calculation of the effect sizes on the same terms for all the studies. Some studies did not report effect sizes, while other did. The way of calculating the effect sizes might slightly differ from one study to another. Therefore, effect sizes for the studies were calculated consistently in order to create comparable sizes to investigate whether and in what domains P4C has the bigger impact.

In previous versions of this review (Ventista, 2018b), the calculation of the effect sizes took into consideration only the post-test performance of the two groups. In this revised version, there is a slightly different calculation of Cohen's d. The calculation of Cohen's *d* when the means and standard deviations of both control and treatment (intervention) group were known were calculated (Cohen, 1988; Morris, 2008).

In order to find the effectiveness of the programme within the groups, a formula which considered both the pre-test and post-test performance of the two groups was used. The equivalence of the two groups in the pre-test could not be reassured, even for randomised controlled trial. As a result, the following formula was used which considered both the pre-test and post-test performance of the two groups.

$$\text{Cohen d} = C_P \times \frac{(Mean\ Treatment_{Posttest} - Mean\ Treatment_{Pretest}) - (Mean\ Comparison_{Posttest} - Mean\ Comparison_{Pretest})}{\sqrt{\frac{(N\ Treatment-1)S^2_{pretest} + (N\ Comparison-1)S^2_{pretest} + (N\ Treatment-1)S^2_{posttest} + (N\ Comparison-1)S^2_{posttest}}{2\,(N_{Treatment} + N_{Comparison} - 2)}}}$$

$$\text{And } C_P = 1 - \frac{3}{4\,(N\ Treatment + N\ comparison - 4) - 1}$$

At the last stage, to respond to the research question the information was combined. Therefore, the effect sizes were categorised according to the skills on which the programme had impact. By investigating the effectiveness of the programme, the literature gaps were also highlighted.

5. Methods: Trial with a Comparison Group (Research Questions 2 and 3)

Having investigated the existing literature, it became apparent that there was a lack of strong evidence regarding the impact of the programme on thinking skills. The second and the third research question of this research examined the P4C impact on critical thinking and creativity. To answer these questions, a comparative evaluation study was conducted. In this chapter, the research design and methods are discussed, and the compromises in the design are justified.

5.1. Research Design

The second and the third research question of this thesis asked whether P4C has an impact on the critical thinking and creativity of pupils who attend primary schools in England. To answer these two research questions, a study with a comparison group was conducted. This research design is recommended to establish the relationship between an intervention and its impact (Cohen, Manion & Morrison, 2007; Shadish, Cook & Campbell, 2002).

To investigate the causal relationship between P4C and thinking skills, an experimental design and specifically a randomised controlled trial could have been a strong design. Consequently, the first target was to achieve random allocation of participants or schools within a control and an experimental group. As it became apparent in the evaluation of research designs in the previous chapter, randomised controlled trials are considered the most robust research design for this purpose. The research design initially involved randomisation within the groups and pre-test and post-test interventions.

The research design could be represented as follows (Gorard, 2013):

R	O_1	X	O_2
R	O_1		O_2

R: random assignment between the groups

$O_{1:}$ Pre-tests: Critical Thinking and Creativity Assessments

X: P4C intervention

$O_{2:}$ Post-tests: Critical Thinking and Creativity Assessments

Nevertheless, it was infeasible to implement this design and randomise within the two groups. This is because P4C training is not freely available and this project was

unfunded. As a next step, I contacted SAPERE and specifically Mr Bob House, and they kindly decided to help me with my research. In other words, the intervention group consisted of schools which had already agreed to receive training by SAPERE. As a result, the training cost would be covered by the schools and provided by SAPERE. One of the positive elements of this design is the training consistency. Even though different trainers organised different training days, all the courses followed the same guidelines by SAPERE. SAPERE allowed me to attend the training courses of schools. This made me aware of the training content. Moreover, SAPERE helped the recruitment process and SAPERE trainers put me in touch with schools. To sum up, the recruitment of the intervention group was facilitated by SAPERE and their help concerning the training provided and the resources was central for the success of this research.

Even though there was no randomisation within the groups, there was an attempt to recruit a matching comparison group. However, the recruitment of a matching comparison group was also infeasible because only a few schools were keen on participating in the research project as a comparison group. Therefore, there was a design of a study with an intervention and a comparison group which was not matched or randomised.

However, the existence of a comparator group is very important for the quality of research conducted to answer a causal question. If there is no comparison group in a school trial, then a positive impact following the intervention is not necessarily caused by the intervention. In other words, assuming that the impact is because of the intervention is an example of a post hoc fallacy. Even though a positive or negative impact might occur after an intervention, it does not necessarily mean that the intervention caused the impact. However, by having a comparison group as a counterfactual provides a plausible comparison to know what would have happened without the intervention (Shadish, Cook & Campbell, 2002), since the comparison group is influenced by the same factors except for the intervention

Therefore, the final research design was a simple two-group study. It started in September 2016 and ended in June 2017. The P4C intervention lasted for ten months. Since the cases involved were not be randomly allocated to groups and there was no matching, in design notation the research design can be presented as:

N O_1 X O_2
N O_1 O_2

$O_{1:}$ Pre-tests: Critical Thinking and Creativity Assessments (September 2016)
X: P4C intervention
$O_{2:}$ Post-tests: Critical Thinking and Creativity Assessments (June 2017)

This design was adopted for comparison of the results of the pre-tests and the post-tests between the experimental and the comparison group. The means of the post-test of the two scores were compared with Cohen d effect size (Cohen, 1988). The uses of effect sizes are discussed later in this chapter.

5.2. School recruitment

School recruitment was one of the biggest challenges of this trial. This led to some compromises in the research design. The adjustments made to the research design also show the dynamic relationship between the research design and recruitment and demonstrate the reasons why a simple two-group controlled trial was conducted.

5.2.1. Intervention group

There were three phases in the recruitment process of the intervention group presented in Table 5.1.

Table 5.1. Intervention Group Recruitment

1st phase	Number of schools receiving P4C training in 2015	48
	Number of schools contacted	47
	Number of schools excluded	1
	Number of schools consented	9
2nd phase	Contact schools which received training before 2015	31
	Number of schools contacted	31
	Number of schools excluded	0
	Number of schools consented	2
3rd phase	Number of teachers approached via training events	More than 10
	Number of cases included	5

	Number of schools consented	1

In the first phase, schools which recently received P4C training by SAPERE were contacted. One school was not contacted to participate in the research because it was a special education school. This research does not exclude all students with special education needs. SEN students who attended a mainstream school were still included in the study. However, it was accepted that different assessments should have been sent in a special education school and there was no time to construct these. I considered it unethical to send the same measurement tools to these students without any adjustment to their needs.

Due to the low number of headteachers who consented during the first phase, the schools which received training before 2015 were also contacted. Even though not included in the initial design of the study, schools with different starting points of implementing P4C enabled the creation of regressions for the relationship between time being involved in P4C sessions and critical thinking and creativity performance.

Finally, in the third phase teachers were contacted during the SAPERE training events. As a researcher I had to attend the SAPERE training, so I will be informed about the specific implementation of P4C and the guidelines suggested by SAPERE. The attendance of the training events though was also considered an opportunity to recruit schools. Concerning the third phase of recruitment, Table 5.1. reports the number of teachers instead of the number of schools because it is common for two or more teachers from the same school to attend the training.

The third phase was not as successful as expected. The teachers who attended the events were rarely Year 5 teachers. Only two of the teachers I met on training days were Year 5 teachers. There were also three teachers willing to pass my contact details to the Year 5 teachers of their schools. As a result, the recruitment of schools via the SAPERE training events was not very successful. The schools which were recruited in the intervention group were 12 in total.

5.2.2. Comparison group

Initially, there was an attempt for recruitment of a matching comparison group. Matching schools were sought in order to be contacted and recruited as a comparison group.

I decided to try to recruit a comparison group matched on shared characteristics, such as proportion of children receiving Free School Meals, proportion of Special Education Needs students and students with English as additional Language and Ofsted reports, accessed via existing datasets. The Department for Education league tables would allow a comparison between the schools based on the performance levels achieved. However, the performance levels do not seem as precise, as the fine scores. Therefore, I decided that the Families of Schools Database provided by Education Endowment Foundation (https://educationendowmentfoundation.org.uk/resources/families-of-schools-database) includes more detailed comparisons. According to this database, the primary schools in England are matched with similar schools based on various criteria and as the website is interactive the criteria can be chosen by the user. When a school and a criterion are chosen, then another school is suggested as the perfect match and a group of schools are recommended as belonging in the same family schools. For each of the intervention schools which consent to participate in the intervention group, the Families of Schools Database suggested more than one schools. These were matched based on the geographic proximity, performance of students (fine score) and attainment gap between premium and non-premium pupils.

It should be highlighted that in fact equal number of schools (more than 80) were approached for the participation in the comparison group (see Table 5.2.). The recruitment of the comparison group was more difficult than I expected and a less robust recruitment process was used. The initial effort to take into consideration only matching criteria was abandoned due to the low consent rate. At the last stage, any school in England not implementing P4C could participate as a comparison group. I tried to look for volunteer schools and at the schools which did not implement P4C and co-operated with the PGCE programme at my own institution were contacted.

Some of the schools which were considered as appropriate for matching in the already recruited intervention group found to implement P4C. These schools showed interest in the research, but they had to be categorised in the intervention group. Consequently, during the attempt to recruit the schools for the comparison group, 6 schools were recruited for the comparison group and 3 additional schools for the intervention group.

Table 5.2. Comparison group Recruitment

Phase	Type of Approach	Number of schools contacted	Number of Schools consented
1st phase	Number of matched schools contacted based on the DfE tables and Families of Schools Database	76	6 schools (3 schools were recruited for the comparison group. The other 3 schools were contacted to be in the comparison group, but they were found to implement P4C and trained by SAPERE. As a result they were included in the intervention group)
2nd phase	Number of schools approached via School of Education at Durham University and volunteering	9	3

Even though the recruitment of matching schools would have provided a robust research design, this recruitment process was not successful. Each of the school to be contacted with two different methods (e-mails and phone calls) and at least two times (for phone guide used for approach of the schools, see Appendix 1a). The school offices were initially conducted via e-mail. I also informed them about a phone call in the following days. The school office usually acted as a gatekeeper explaining that the school was too busy to participate in a research and rarely passing the calls to the school teachers or headteachers and the e-mails for expression of interest were rarely replied to.

It is likely that this recruitment obstacle is associated with the lack of extrinsic motivation and research funding. If they were resources or training offered to the control schools with a waiting list design, the rate of consent might have been higher. In other words, there was no incentive for control schools except for school reports to be sent at the end of the study.

The intervention group schools having already taken or agreed to take SAPERE training had a motivation to investigate the effectiveness of P4C programme in their

students. However, the schools in the comparison group were probably less interested in the P4C effectiveness. If there was funding, a crossover research design would have adopted. According to this design, the comparison group gets the intervention and a second post-test follows (Shadish, Cook & Campbell, 2002, p.268). However, due to the lack of funding, the schools in the comparison group were not given any reward for their participation. Consequently, the difficulty in recruitment of comparison group led to compromises in the research design.

5.3. Sample

The two research questions to be addressed via the study were whether critical thinking and creativity can be developed when P4C is implemented in primary schools in England. Therefore, the targeted sample was students attending primary schools in England. This project targeted only a particular age group of primary school students because it would have been difficult to find or construct age appropriate assessments for both constructs for all the students in primary schools.

Specifically, Year 5 classrooms were invited to participate. The reason why Year 5 students were invited to participate was mainly because of their reading ability. Given that the students were required to sit a critical thinking assessment, which involved thinking problems, the students should have had a sufficient level of reading skill and comprehension. Moreover, the creativity test used would require the students to produce responses in a written form. Therefore, I chose a year group with developed basic literacy skills which would not impede them from performing in the assessments. Students who attended Year 6 would possibly be more suitable because there is an available critical thinking assessment which could be used for Year 6 students (Ennis & Weir, 1985). However, it was judged that Year 6 classrooms would be less likely to participate - particularly in the post-test at the end of the school year - because of accountability reasons at the end of Key Stage 2.

The initial sampling method I aimed to use was random from the population of schools in England. However, as it has already been explained, the lack of funding did not enable me to adopt this design. The treatment group was contacted via SAPERE. Therefore, the sampling method is what is called convenience sampling (Cohen, Manion, & Morrison, 2007) since it involved schools which could be easily accessed. Similarly, the schools in the comparison group were chosen based on convenience sampling. Therefore, no generalisation claims are made in the results section since this

sample is not representative of a particular population. However, initial differences are handled by taking the pre-test into account.

Furthermore, the sampling and participation in the study took place on a school level, and not on an individual or classroom level (Figure 5.1.). It is also important to examine the characteristics of the schools which participated in the study. These are presented in Table 5.3 and they are based on Edubase of Department for Education (n.d.) and Performance Tables (GOV.UK, n.d.) for the academic year 2016-2017. This is the school year when the evaluation study was conducted. In this table, the schools in the intervention group and the comparison group are mixed. This is because they are presented with the codes used for the blind marking of the assessments. Schools were mixed for a blind marking in order to reduce grading bias.

It becomes apparent that schools participated in the study came from different parts of the country (Table 5.3). The majority of the schools in the comparison group came from the North. This is due to the fact that the second phase of recruitment process of the comparison group took place via the School of Education at Durham University and therefore schools nearby agreed to participate.

Only two comparison schools had a proportion of SEN students more than the national average. The fact that P4C schools had low proportion of SEN students (Table 5.4) might indicate that schools with higher proportion of SEN students introduce different interventions in their schools instead of P4C. Concerning the OFSTED rating, there was no particular difference between the comparison and the intervention schools (Table 5.3). The proportion of students with EAL varied based on location and three comparison schools located in the North had a few students with EAL (Table 5.3). Hence, the average of EAL students in the comparison group was significantly lower to this of the intervention group or the national average (Table 5.4).

Figure 5.1. Participant Flow Chart for Research Questions 2 and 3.

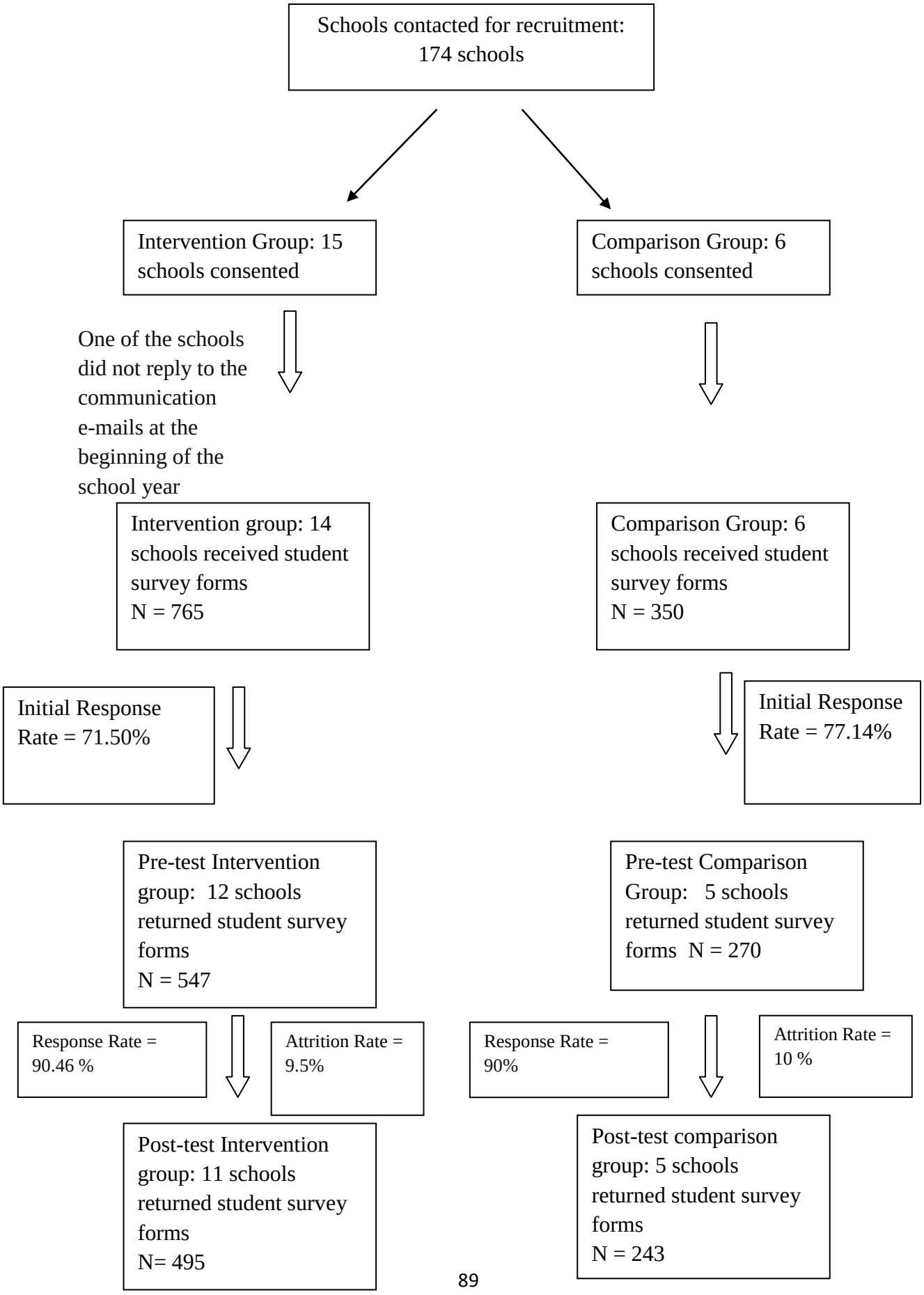


Table 5.3. Participating School Characteristics for the Year 2016-2017.

	Group in the study	Type of school	Area	N pupils	Gender	SEN students in the school and EHC plan	FSM students in the school during the last 6 years	EAL students in the school	Ofsted Rating
1	Intervention	Academy Converter	South West	435	Mixed	0.2%	11.1%	0.3%	Good
2	Intervention	Community School	North	378	Mixed	1.3%	48.9%	21.3%	Good
3	Intervention	Community School	East Midlands	282	Mixed	0.4%	30.5%	3.7%	Requires Improvement
4	Comparison	Academy Sponsor Led	North East	2,561	Mixed	5%	47.1%	1.1%	Requires Improvement
5	Intervention	Academy Converter	East Midlands	349	Mixed	0.6%	16.3%	2.6%	Good
6	Intervention	Community School	North East	141	Mixed	0.7%	40.4%	2.1%	Good
7	Comparison	Voluntary Aided School	North East	258	Mixed	3.1%	18%	3.4%	Outstanding
8	Comparison	Community School	South East	532	Mixed	0.9%	5.1%	27%	Good
9	Intervention	Academy Converter	South East	362	Mixed	1.1%	19.1%	46.1%	Good
10	Intervention	Community School	East	439	Mixed	2.5%	22.7%	65.4%	Good
11	Intervention	Academy Converter	North East	506	Mixed	0.6%	38.5%	2.5%	Outstanding
12	Intervention	Academy Converter	South West	422	Mixed	1.4%	9.2%	2.5%	Good
13	Comparison	Academy Sponsor Led	East	164	Mixed	0.6%	40.9%	25.2%	Good
14	Comparison	Academy Converter	North	252	Mixed	2%	27.7%	0%	Good
15	Intervention	Academy Converter	North	347	Mixed	0.9%	28.2%	10.1%	Good
16	Intervention	Academy Sponsor Led	London	534	Mixed	0.2%	38.8%	69.6%	Outstanding
17	Intervention	Community School	East	456	Mixed	0.9%	20.1%	65.4%	Good

Table 5.4. Comparison of pupils' characteristics in the two groups with the National Averages for the Year 2016-2017.

Categories	National Average	Intervention Group Average	Intervention Group N of students in the category/ total N in the group	Comparison Group Average	Comparison Group N of students in the category/ total N in the group
Special educational needs (SEN) or education, health and care (EHC) plan for 2016-2017	2.9 %	0.9%	42 / 4,651	3.9%	145 / 3,767
FSM at any time during the past 6 years	24.9 %	26.5%	1,233/4,651	37.6%	1,416/3,767
Percentage for pupils whose first language is not English (EAL)	20.8 %	27.7%	1,287/4,651	5.9%	222/3,767

5.4. Response Rate and Missing Data

In this section, the response rate and the missing bias occurred during the study are reported (Figure 5.1). There are two different response rates reported. The first response rate refers to the number of pre-test forms sent and then returned whilst the second one to the drop out from the pre-test to the post-test.

The initial response rate for the intervention and the comparison group were 72% and 77% respectively. I would like to argue that the initial response rate is probably higher than the reported one. This is because the numbers of forms sent to the schools were the number of forms that the headteachers asked for the Year 5 students. All the headteachers asked for forms in a round number (e.g. they asked 50 or 60 forms). It is probably improbable that all the schools had a round number of Year 5

students. This suggests that the school leaders probably did not ask forms for the precise number of students, but instead they asked for more forms than the actual number of students.

After that initial response, the same schools were asked to complete assessments at the end of the school year. At this stage attrition bias took place. This bias occurs due to the fact that participants drop out (Gorard, 2013; Torgerson & Torgerson, 2008). Missing data are reported. Due to anonymity of questionnaires, it was not feasible to match pre-test and post-test performance. Therefore, the cases with pre-test data, but post-data missing could not have been excluded or been treated differently. However, as I will explain later in this chapter, the number of cases need to disturb the finding (NNTD) was calculated. This is a way used to estimate the number of missing cases were adequate to change the findings of this thesis.

5.5. Teacher training

The intervention schools were all trained by SAPERE. As it was discussed in the second chapter of this thesis, SAPERE is a UK charity registered in England, it is inspired by the ideas of Matthew Lipman and provides P4C training to practitioners. All the intervention schools of this study received SAPERE training. Even though SAPERE is the most well-known P4C trainer, P4C training is not offered exclusively only by SAPERE in England.

For the purposes of this study as a researcher I was not obliged to have P4C training, since I was the programme evaluator and I was not a P4C teacher. Nevertheless, I considered it important to find out the type of training that the intervention schools received and as a result I attended SAPERE training. SAPERE training is provided by different trainers but there is consistency among the training each of them provides.

SAPERE offers three different levels of training. Each level has a pre-requisite the previous levels. I only attended the Levels 1 and 2 because there was no intervention school in my study known to have received Level 3 training. The training course material of SAPERE states Level 1 includes only an introduction at P4C, whilst Level 2 includes two different sub-levels. Level 2a involves the improvement of P4C session and it emphasises the development of the 4 Cs (Creative, Critical, Caring and Collaborative). The link between the concepts of this thesis and two of the 4 Cs is

apparent. Level 2b involves topics such as how P4C improves the school ethos and political and ethical implications of P4C.

There is a gradual building of knowledge with P4C training and most of the interventions schools of the study received only Level 1 training. This study required 'clean' schools which had not previously received the intervention which might affect their pre-test performance.

5.6. Measurement Tools

The research design required the participants to be assessed twice. Their performance in critical thinking and creativity was evaluated at the beginning and the end of the school year. The participants of both the intervention and the comparison group were assessed. The performance of the intervention group was compared to the comparison group in order to investigate the impact that P4C had on the development of their critical thinking and creativity. For the purposes of this research, a new assessment tool was created to assess critical thinking and creativity. In Chapter 7 there is a detailed description of the assessments used as pre-test and post-test. The methods of scoring are presented in Chapter 8. Finally, since they are new assessments created for the purposes of this research, they were piloted, and the results of this piloting will be presented in Chapter 9.

5.7. Fidelity to implementation

There were two different measurements to examine the implementation fidelity. First, as a researcher I visited schools from the intervention and the comparison group in order to examine whether they implemented the intervention as they were trained. Furthermore, observation of the comparison group facilitated the examination of whether they did not implement the intervention and investigate the practice they implemented instead of P4C. After the attendance of classes in the schools, there was usually a discussion with the classroom teacher about the fidelity to implementation. The school visits provided with in-depth data about the fidelity to implementation. However, as it has already been acknowledged from previous studies, school observations are expensive and the researcher cannot be constantly present. It can be questioned to what extent the teachers behave in the same way when they are observed (Topping, 2018).

As a second way to examine implementation fidelity, schools were asked to report the regularity of implementing P4C during that academic year. This was to estimate fidelity to implementation. In order to collect this information, a questionnaire was sent to the schools to be completed by the classroom teachers. The questions were open-ended in order not to lead the teacher to choose a particular response. One P4C school reported that they stopped implementing P4C during the year, whilst one classroom in the comparison group started implementing P4C during the academic year. Therefore, the intention to the treat analysis definitely included some error in this study.

The fidelity can be a moderator factor of whether an intervention succeeds or not and in the case of this study it is a factor which could not be controlled. This frequency of implementation is reported in Table 5.5. This variable is included in a regression for critical thinking development in Chapter 11. It was examined whether this variable can be a predictor of development of critical thinking.

Table 5.5. Frequency of Treatment in Intervention Schools.

Regularity of P4C Sessions in the Intervention schools	**Number of Schools**
2 30-minutes session in a week (twice a week)	1
Weekly	7
Twice a month	1
Once a month	1
Stopped implementing	1
Total	11

5.8. Treatment Diffusion

Calsyn (2000) emphasised the fact that treatment diffusion is a crucial factor to be examined. Treatment diffusion refers to the amount of intervention which was received by the comparison group. This can be a threat to validity of the study, reducing the apparent effect size. If the comparison group also started receiving the intervention, in the post-test results might show that the intervention did not have an impact.

Treatment diffusion usually takes place when the comparison group learns about the intervention used in the intervention group and imitates the intervention. Diffusion was not judged a big threat to the validity of the study. It is apparent that

treatment diffusion is more likely to occur when participants from the two groups interact with each other. The schools in this research project were located across England and there was no direct communication between the participants in the comparison and intervention group. This means that the intervention group was not expected to influence the comparison group. To examine for treatment diffusion, I also visited comparison schools in order to examine whether they implement P4C.

During the visits of the schools in the comparison group, there was an attempt to identify whether P4C started being implemented in the school. In one of the comparison schools (school code 4), it was found that one of the Year 5 classes started implementing P4C sessions. This was not included in the analysis, because the analysis adopted was 'intention to treat'.

5.9. Intention to treat

It was not feasible to verify whether teachers' answers about the regularity of the sessions in the intervention group were accurate. Even though teachers were asked to report the regularity of the sessions in forms sent with the assessments at the beginning and the end of the school year, this did not mean that the teachers implemented P4C as regularly as they reported. Likewise, teachers in the comparison group might not have reported P4C training or P4C sessions, but they might have started doing it or attended a training session after the allocation in the comparison group.

Thus, the results of this study were analysed according to the intention to treat. Intention-to-treat is implemented in trials to face the problem of non-compliance and missing outcomes and it ignores everything that happens after randomisation (Gupta, 2011). In the case of this study, withdrawal and noncompliance that happened after allocation in the two groups were ignored. It was assumed that Year 5 students in the intervention group had P4C sessions, which were implemented according to the methods suggested in SAPERE training. Also, it was assumed that Year 5 students in the comparison group did not receive P4C.

5.10. Analysis: Effect Sizes

In order to respond to the research questions and identify whether P4C has an impact on critical thinking and creativity, Cohen's d effect sizes were calculated (Cohen, 1988). The following effect size was used to calculate effect sizes within the same group for pre-test and post-test as presented in section 4.4. of this thesis.

The analysis was pre-decided to avoid cherry-picking. Significance testing was avoided. In order for significance testing to be used, there should be random sampling, random allocation within the groups and no dropout in the study (Gorard, 2016). It is apparent that none of these applies to the design of this research and therefore significance testing could not have been implemented.

To express the extent to which the research findings are trustworthy, Gorard and Gorard (2016) introduced the number of cases need to disturb the finding (NNTD). In other words, they asked how many counterfactual cases were needed to change the finding of the study, if these cases demonstrated 'opposite' performance to the findings by a standard deviation each. If the number of cases needed to disturb a finding is a small one, then the finding is not strong, whilst if this number is big then the finding is trustworthy.

Gorard and Gorard (2016) suggested an iterative way of calculating this number, which was later simplified by Kuha and Sturgis (2016). Gorard, See and Siddiqui (2017) developed this way of calculating of the number of counter-factual cases to disturb a finding. In this study, this number is also calculated and accompanies the effect sizes, in order to demonstrate the robustness of the study.

$$NNTD = |effect\ size| \times N\ of\ cases\ in\ the\ smaller group\ in\ comparison$$

5.11. Regression

Predictive models for the performance of the critical thinking and creativity were created. These were based on specific variables provided by the students and examined how much of the variance of their performance on the post-test can be explained by these variables.

The variables included in the model were separated in two steps of regression. The first step included students' characteristics and initial performance. For the gender, boys were coded with the number 1 and girls with 2. Age was considered as reported by the students. Pre-test performance was not very accurate, because it was not based on the performance of the individual in the assessment. Particularly, the individual pre-test performance was not known due to anonymity of questionnaires. Therefore, the mean score of all the students of each school was used as the baseline performance for all the students in the school. This approach cannot account for cases appearing only in the pre- or post-test.

In the second step, there were three variables related to P4C. The intention to treat examined whether the students were in the comparison group or intervention group based on the initial separation in two groups (comparison group coded as 0 and intervention group as 1). The frequency of the sessions and the years of implementation were based on the reporting of the teachers in the teacher questionnaires sent. The variable with the frequency was coded as 0 for no implementation, 1 for monthly implementation, 2 for twice a month, 4 for weekly implementation or greater implementation.

5.12. Ethics

The research adhered to Durham University data security (for ethics approval letter, see appendix) and BERA ethical guidelines (BERA, 2011). The headteacher or the classroom teacher was initially contacted. In the consent forms, the schools were fully informed about the research, its purposes, its benefits and consequences and their role in the research. Until written consent was obtained, no surveys were sent to the school. The students and their teachers were informed of the right to withdraw from the study at any point.

All participants in the primary data collection were asked for informed consent and their right to withdraw from the study was made explicit to them. The questionnaires of critical thinking and creativity were anonymous and since the researcher was not present in the schools when the data collection and assessments took place, no individual is identifiable.

There is also another ethical issue which was considered. It might be claimed that the comparison group has the right to be treated equally and it was unethical for this group not to receive the intervention. However, Gorard (2013) argued that people receive and not receive interventions all the time, and particularly in the case of research trials until the intervention is proven to be advantageous, the comparison group cannot be considered deprived. The schools in the comparison group had no intention of conducting P4C and they were not stopped by proceeding in the implementation.

5.13. Important Dates

Time was an important factor for the trial and therefore important dates should be presented. For a better understanding of the timeline of the project, Table 5.6 presents the key dates.

Table 5.6. Key Dates for the Project

Date	Event
2nd March 2016	Obtained ethical approval for my study.
March 2016	I contacted SAPERE. Intervention group recruitment – Phases 1 and 2.
April 2016	Attended SAPERE training (Level 2b) at Hull. Intervention group recruitment – Phase 3.
April 2016	Creation of YouTube video 'Philosophy for Children: Get Involved' for the link sent in the school communication e-mails https://www.youtube.com/watch?v=MULaNDH-PuI
End of May 2016	Measurement tools piloting in a school. This school was not included in the comparative evaluation study.
June 2016	Attended SAPERE training (Level 1) at Newcastle. Intervention group recruitment - Phase 3
June 2016 - September 2016	Recruitment of comparison group
September 2016	Pre-test
October 2016 - May 2017	P4C Intervention Visit schools for observations. Informal interviews with teachers.
June 2017	Post-tests

6. Methods: Secondary Data Analysis (Research Question 4)

6.1. Research Design

The fourth research question of this thesis examined the impact of P4C on attainment. To answer this question, this thesis used secondary data analysis of pupil-level data from Department for Education. Smith (2017) argued that the secondary data analysis is a democratic research method, because it enables all researchers to be engaged in research. Fieldwork requires time and money. When secondary data analysis is used, research is not a privilege of these researchers who have funding or time to conduct research. Concerning educational research, she recommended the Census Data for secondary data analysis in the UK. This is the data used by this thesis and particularly the National Pupil Database (NPD). However, the process from the time of first application to data receipt lasted approximately one year. The application required several revisions. Therefore, I argue that even though secondary data analysis is democratic process, this type of project is not always faster conducted than a fieldwork project.

However, using NPD gave me access to more data than I could expect to collect via fieldwork. The population for this research question was all students attending state-funded primary schools in England. Private schools were excluded from this analysis to avoid over-complicating it. The 'experimental' group included the schools that implement P4C for the same cohort of students from Year 3 to Year 6. The experimental group included all the students who attended schools which received P4C training from 2010-2011. In the comparison group there were all the students who attended schools which have never been trained in P4C by SAPERE since they were not included in SAPERE database with schools which received training.

Searching for the impact of P4C on attainment, a longitudinal experimental design is a strong design, because both experimental and longitudinal studies are appropriate designs to establish causal relationships (Cohen, Manion, & Morrison, 2007). Key Stage 1 (Year 2) performance between control and intervention group were compared to the subsequent Key Stage 2 (Year 6) performance of the same cohort of students.

To be precise, the Key Stage 1 results used as a baseline assessment from the National Pupil Database included the variables Key Stage 1 Reading Points, Key Stage 1 Writing Points and Key Stage 1 Maths Points. As a post-test, the Key Stage 2 results

were used. The aforementioned variables were compared to the Reading Fine Score, Grammar, Punctuation and Spelling (GPS) Fine Score and Maths Fine Score.

Specifically, the comparison included:

N	O_1	(X)	O_2
N	O_1		O_2

$O_{1:}$ Key Stage 1 results 2011 (3 subjects)
(X): Naturally occurring P4C intervention (2011-2015)
$O_{2:}$ Key Stage 2 results 2015 (3 subjects)

6.2. Cases

SAPERE provided me with a list of schools which have been registered for P4C training and the date of training. I investigated this research question based on intention to treat, which means that I assumed after the training the schools implement of P4C. The intervention group was all schools which received P4C training from 2010-2011 and no later. All the schools which were not in the list were allocated to the comparison group. The schools which did not receive SAPERE training at any point after 2010 were included in the comparison group (and this could only reduce the effect size).

SAPERE lists included infant schools and secondary schools, such as grammar schools. Nevertheless, this analysis focused in the progress of the schools during Key Stage 2 and therefore infant schools and secondary schools were excluded. Moreover, in the SAPERE lists there were also Welsh establishments and a few international schools. The cases in this analysis were schools in England and therefore these schools were not included.

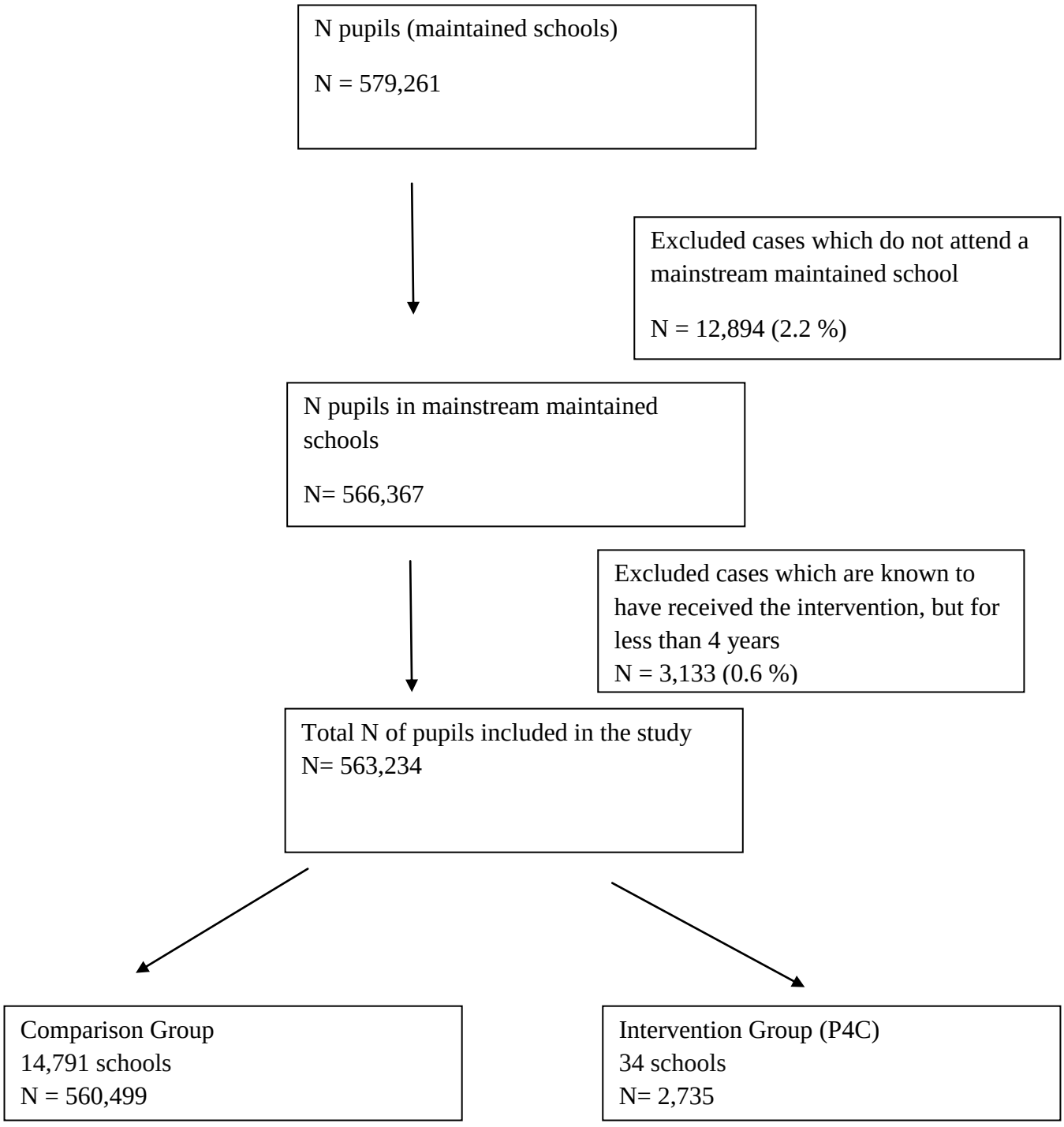
Since in the analysis there were only mainstream maintained schools in England, then the option ‘select cases’ were used in SPSS. Therefore, select cases which satisfied the condition of being mainstream maintained schools (including academies) were included. By using the National Pupil Database, the variable KS2_MMSCH was used and only the schools coded with 1 (=yes) were in the analysis.

From the EduBase I found the URN for each of the schools and I included these as P4C schools these establishments. Consequently, from the 48 establishments provided by SAPERE, only 34 schools were included in the analysis. This is because the URNs from the others were not in the National Pupil Database, since they were

infant school, secondary school or Welsh establishment. A new variable (P4C) was created in the database. These 34 schools were coded as 1, whilst others were coded as non-P4C schools with code 0.

Schools which are known to have received training from 2012-2015 were excluded from the analysis because they did not belong to either comparison or intervention group. They were not included in the comparison group because that would have reduced the size of the effect sizes, since the students who took Key Stage 2 assessments would have had received P4C for a few years before 2015. Furthermore, they were not included in the intervention group because the students did not receive P4C for four years successively and this comparison aimed to examine the impact of P4C when it is implemented longitudinally (code 3 in the variable P4C – these schools were excluded from the analysis). The process and the number of cases included and excluded are presented in Figure 6.1.

Figure 6.1. Participant Flow Chart for Research Question 4.



6.3. Missing Data

Missing data are not considered random and therefore the missing data were examined. The analysis also considered the eligibility of students for Free School Meals the last six years as an indicator of disadvantage. There was no missing data on Ever FSM6 in NPD (see Table 6.1. for descriptive statistics).

Table 6.1. Frequency of students based on their EverFSM6 Meals Eligibility.

	Frequency	Percent
Non-FSM	389,415	69.1 %
FSM	173,819	30.9 %
Total	563,234	100 %

However, there was missing data when students' assessments were considered. The numbers of students whose data is provided and whose data is missing are presented in Table 6.2. More missing data are observed for Key Stage 1 results compared to Key Stage 2 (see Table 6.2.). However, even for the Key Stage 1 results the missing data are less than 5 %, whilst for the Key Stage 2 results the missing data are less than 1%.

Table 6.2. Frequency of Pupils' Assessment Data in Maintenance Schools.

Type of Assessment	Total N	Total N of students with valid data (no missing)	N of students whose data is Missing
Reading Key Stage 1 (2011)	563,234	536,540	26,694
Writing Key Stage 1` (2011)	563,234	536,526	26,708
Maths Key Stage 1 (2011)	563,234	536,423	26,811
Reading Key Stage 2 (2015)	563,234	561,639	1,595
GPS Key Stage 2 (2015)	563,234	561,631	1,603
Maths Key	563,234	560,998	2,236

Stage 2 (2015)			

Cases with missing data could be excluded from the analysis. However, it was not desirable to exclude these cases completely from the analysis. This is also due to the fact that they are not the same cases whose data is missing from the assessments at Key Stage 1 and 2 at the same subject. Hence, there would be a high number of cases that I would have to exclude which could reach up to 5% of the sample. In order not to exclude these cases, the missing data from each assessment were replaced with the mean score that students scored in that assessment (see Table 6.3. for the mean score per subject used to replace the missing data).

Table 6.3. Mean Score per subject.

Year Group	Variable	Mean
Key Stage 1	Writing Points	14.53
Key Stage 1	Reading Points	15.88
Key Stage 1	Maths Points	15.83
Key Stage 2	GPS Fine	4.85
Key Stage 2	Reading Fine	4.78
Key Stage 2	Maths Fine	4.85

6.4. Analysis

A few points should be clarified in relation to the analysis implemented. The impact on reading, writing and Maths was examined. However, Key Stage 1 and Key Stage 2 assessments use different scales. In order to compare, the Key Stage 1 and Key Stage 2 results, z-scores were used. Finally, these z-scores were used in order to calculate the effect sizes for these three areas. The effect sizes considered both pre-test and post-test performance.

Furthermore, a previous study examined the P4C impact on attainment (Gorard, Siddiqui & See, 2017) found higher impact on disadvantaged students eligible for Free School Meals (FSM). To examine whether this could be the case when there is a long term implementation of P4C, this thesis examined separately the P4C impact on students eligible to FSM. Particularly, the variable of whether the pupils are known to be eligible for Free School Meals the last six years was used (see Table 6.4). I considered the indicator for FSM for the last six years, because that indicator fitted better with the longitudinal approach of this study.

Table 6.4. Number of FSM or non-FSM students per group.

	Ever FSM6 (percent of students from the overall group)	Non-FSM (percent of the students from the overall group)
Comparison Group	173,020 (69.1%)	387,479 (30.9%)
Intervention Group	799 (29.2%)	1,936 (70.8%)

Finally, the analysis which took place was intention-to-treat as in the previous two research questions. The schools which received training before 2012 were included in the intervention group. There was the assumption that the schools which received training before 2012 kept implementing P4C during these four years. Similarly, schools not included in SAPERE lists were considered as non-receiving P4C.

7. Measurement Tools for Research Questions 2 and 3

This chapter is related to the second and third research questions of this thesis, which focus on P4C impact on critical thinking and creativity. Particularly, this chapter discusses the assessments used for the evaluation of critical thinking and creativity. A whole chapter discusses the particular issue, because it might be debatable to what extent these two skills can be evaluated effectively. Therefore, it is crucial to explain explicitly all the informed decisions concerning the selection and construction of items in the assessments. For example, this chapter explains the reason why the assessment of critical thinking includes more items compared to the creativity one.

Initially the choice of particular assessments is justified. For the purposes of this research, creativity assessment was a combination of existing measurement tools, whilst the critical thinking assessment was created for the purposes of this research. Then, decisions of the implementation of assessment are presented. The parallel forms used are also investigated.

Having discussed the construction of the two forms as a whole, an in-depth explanation of the rationale of each item included in the assessments follows. Particularly, the inclusion of each item in the assessments is justified. Furthermore, there will be a clear matching between the working definitions of the two constructs as discussed in Chapter 3 to the items of these assessments. Hence, the design of the thinking assessments is thoroughly discussed.

Finally, considering the significance of the assessments for the validity of the study, the measurement tools are evaluated. The psychometric properties of the assessments are explored and how the assessments satisfy the criteria of effective assessment suggested by Fisher and Scriven (1997).

7.1. The choice of measurement tools

There are many assessment purposes (Newton, 2007). For this research, assessments were selected in order to fit the purpose. The purpose of the assessments was to evaluate critical thinking and creativity in order to examine the impact that P4C intervention has on them. Therefore, the assessments did not promote Assessment for Learning and they did not aim to identify students' misconceptions or weaknesses in order to support further learning. The purpose of the assessment was to be sensitive to

measure the performance of cohorts in critical thinking and creativity tasks for the production of scores which enabled comparisons of group performances.

The choice of the measurement tools was a long process. I started researching about the measurement tools of critical thinking and creativity by piloting existing measurement tools in secondary schools in England and Greece (Ventista, 2018a; Ventista & Coe, 2015). Some of the findings of that research were useful guides for the assessments in this research. These findings can be summarised:

- Assessment tools of creativity and critical thinking which evaluate these two constructs as general skills can be reliable and valid.
- Existing creativity tools appear to be reliable concerning their internal consistency. Moreover, they are characterised by criterion, convergent and discriminant validity.
- Critical thinking items can be knowledge or culture dependent.
- The context used for the items of critical thinking problems should not be closely related to the daily experience of the students. For example, names of towns they visited should not be used. This is due to the fact that this context could lead to stimulus adherence and the students might not be as critical as the test demands.

All the above findings were considered for the selection and the construction of the assessments for this trial. Therefore, I could proceed in this project and have some confidence that I could measure critical thinking and creativity independently of a skill for Year 5 students. The piloting of the creativity assessments in my previous study provided positive results (Ventista, 2018a) and the assessment was judged age appropriate for Year 5 students. Therefore, the creativity assessment for this study was based on a combination of existing assessments.

Concerning the critical thinking assessments, there was a wide literature review to identify an assessment appropriate for Year 5 students. However, there was no appropriate existing assessment found in the bibliography. Ennis and Weir (1985) claimed that their assessment can be used for Year 6 students or older. However, when I piloted it with secondary school students (Ventista, 2018a), I found that it was knowledge dependent. Some of its items required particular knowledge and additionally the reading of the essay required high level of literacy. Thus, its use was

avoided. The authors of the Halpern critical thinking assessment (2010) and the Watson and Glaser critical thinking assessment (2002) did not recommend their assessments for young students, so these assessments were not age appropriate for the Year 5 participants in this study.

Another example of a test which was not judged appropriate was appraising observations (Norris & King, 1984). This test did not cover adequately all the aspects included in the working definition of critical thinking of this thesis, since it focused only on particular aspects. The New Jersey Test of Reasoning Skills (Shipman, 1983) was also rejected because it is the test which was commonly used in P4C studies and I wanted to separate the P4C tradition from this study and proceed to an independent evaluator. Moreover, the test is relatively lengthy which is not appropriate for the students of this age. There were other assessments which were rejected because of length, lack of age appropriateness and lack of link to the definition of critical thinking accepted by this thesis.

The Cornell Test Level X (Ennis, Millman & Tomko, 2005) was the only test which was judged age appropriate by the authors of the assessment. However, the test requires the reading of long texts. Thus, its use was avoided for reasons of validity (to avoid construct irrelevance), as will be explained later in the psychometric properties of the test. Therefore, there was no measurement tool available for the purposes of this research. The critical thinking assessment was inspired by the Cornell Test Level X, since it is also co-authored by Ennis, whose ideas inspired also the working definition of critical thinking for this research. Nevertheless, a new test was designed specifically for the purposes of this research and particularly a test which could fit the purpose, be age appropriate for Year 5 participants and correspond to the working definition adopted by this research.

7.2. The design and implementation of a unified assessment

In this section, there is a discussion of some main issues concerning the construction and the implementation of the assessments. The initial issues refer to decisions concerning the implementation of the assessment. First, the sequence of the implementation of the assessments will be defended. Then, the reason why this research evaluated critical thinking and creativity with one unified measurement tool instead of two separate tests will be explained. The reasons why the critical thinking assessment includes more questions than the creativity assessment will also be

explained. After that, the time given to the students for the completion of the assessments will be justified. Finally, the collection of additional demographic data will be presented.

7.2.1. The sequence of implementation of assessments

Initially, I decided on the sequence of implementation, the number of tests administered, and the number of items included in the test. Concerning the sequence of assessments, which of the two constructs should be assessed initially and what should follow was to some extent arbitrary. However, I decided on a creativity test followed by a critical thinking test because of the item format. I judged that it would have been better to start with the open-ended creativity questions and then 'restrict' the students' thinking in multiple-choice items for the critical thinking test. Based on my experience of working with students as a teacher, I speculated that some students would have been tired and less motivated to respond to open-ended questions if they were presented at the end of the assessment.

7.2.2. One unified assessment

After deciding on the sequence of the tests, I decided whether the measurement would be made by the implementation of two separate tests or a unified test which evaluates both constructs. The administration of two different tests can reassure that the appropriate assessment timing was kept for both of them. By saying this I mean that having two independent measurement tools was the optimal way to guarantee that each assessment is completed in the provisional time. On the contrary, by administering one test it would have been uncertain how much time the students spent on the open-ended questions and what was the precise time that they spent thinking about the critical thinking problems. As a result the students might have concentrated on one type of questions and they might have lacked time to reply to the questions. This inability to organise the assessment time and split it wisely between the two assessments could lead to poor performance to one of the two tests. Poor performance in one of the two tests due to inability to organise the assessment time does not mean that the students deserve to be judged as un-creative or a-critical. Therefore, the administration of two separate tests could have excluded the possibility of students performing poorly in one construct due to their inability to ensure sufficient time for the items of this construct.

Even though the option of implementing two separate tests seemed to be advantageous to keep the provisional assessment duration, this option had a main

disadvantage. This option appeared to be less pragmatic. The administration of two tests would demand more effort and time. Having decided from the beginning that I was not going to be present during the administration, it sounds reasonable to require the class teachers to administer one instead of two assessments.

7.2.3. The length of critical thinking assessment

The critical thinking test has more items than creativity test. The inclusion of more items does not imply that critical thinking has more facets than creativity or that the measurement of creativity is underestimated compared to critical thinking. The only reason why critical thinking test is longer and includes more items compared to creativity test is because each facet of critical thinking has to be evaluated separately. Each of the items which evaluate creativity can measure more than one aspect of creativity. For example, item one measures fluency, flexibility and innovation. Critical thinking items had to be extremely complicated to evaluate more than one aspect. Therefore, each item evaluates only one facet of the construct and this is the reason why the critical thinking assessment needed more items.

7.2.4. Assessment time

It has been recommended that half an hour is a reasonable amount of time to retain the motivation of primary-school students (Gronlund, 1982, p.32). Hence, the number of items included in this test corresponds to the demand of keeping the duration of assessment within this limit. The multiple-choice critical thinking items of this test are considered more time-consuming compared to multiple-choice items of another type of test. In another type of test, such as a history test, the students are usually required to recall information they memorised and they can immediately search for the correct answer between the options given. In critical thinking multiple-choice items, the students have to spend time thinking. What is more, in some cases when the item demands the students choose the best answer amongst a series of correct answers, the students are required to examine the feasibility of each choice separately and choose the best. Therefore, a relatively small number of items can be found in the assessment used in order to retain the length of the assessment at an age appropriate level.

7.2.5. Demographic characteristics: gender

Except for creativity and critical thinking, the assessments did not include much additional data collection because it was unnecessary for the assessment purposes.

Furthermore, there was an intention for the data to remain anonymous. Even though it would have been probably better to know the names of the students in order to match individual performances and calculate gain scores between pre-test and post-test, I recognised that this element would discourage some schools from consenting to the study.

In the primary data collection, the demographic characteristics collected by the students were very few and were asked at the beginning of the test. The form the students completed had the name of the school written on it. This was necessary for research purposes since the separation of the students in the control and intervention group was school-based and the participants were identified as a member of one of the two groups. Moreover, the students were asked to write a number for their age to examine whether the assessments favoured older students.

The last characteristic collected was the gender. There is a study which identified gender differences in creativity assessments based on the provided stimuli (Kaufman, 2006). However, the creativity assessment was based on existing assessments and it used both a verbal and a non-verbal stimulus. Furthermore, there are studies which include evidence that the format of multiple-choice questions itself is favouring boys (Beller & Gafni, 2000). A recent study from Stanford University (Reardon et al., 2018) revealed that boys perform higher than girls in assessments on multiple-choice questions and test format plays a role in students' performance. Therefore, multiple-choice questions can be biased in favour of boys.

Nevertheless, it is assumed that whatever impact the gender bias had on one group, it would also have had on the other group. This is due to the fact that both groups consisted of a similar percentage of boys and girls (see Table 7.1). In other words, bias might have mattered if there were differences in the proportion that each gender appeared in each of the two groups. Since this is not the case, if there is any gender bias, it affects equally the performance of both groups.

Table 7.1. Gender of Students in each Group

Group	**Pre-test**			**Post-test**		
	Boys	Girls	Blank	Boys	Girls	Blank
Intervention	276	270	1	223	205	0
Comparison	140	130	0	154	156	0

I collected information about gender for the regression included in the results and I thought that this information could be collected in a multiple-choice question. It would be recommended for students of this age not to be required to write much information. Thus, I thought I would include the options 'boy' and 'girl' with a box next to them for the student to tick the option.

However, gender is not such a simple concept. According to the Office for National Statistics (n.d.) in the UK: 'Gender identity is a personal internal perception of oneself, and as such, the gender category with which a person identifies may not match the sex they were assigned at birth'. Thus, gender identity is self-defined. There is a rationale for the construction of items for gender and different policies are followed in different countries (Statistics New Zealand, 2014; 2015). Thus, in the category gender identity there should not be offered only a binary reply. Other options, such as 'agender', 'gender fluid', 'gender queer', 'transmale', 'transfemale', should be included (Treharne & Beres, 2016).

Without aiming to discriminate any student or making them feel uncomfortable, I accepted that offering all these options in a questionnaire might have resulted in questioning by Year 5 students. At the same time, offering the options 'boy, 'girl' and 'other' could equally stimulate the curiosity of Year 5 students. In this case, their teachers would have been asked for explanations which they might have not comfortable to provide. Thus, it is highly likely that the teachers could have been demotivated in distributing the assessments in their classroom to avoid such disturbance. Therefore, the question about the gender was included as open-ended. It would be preferable to ask students to write a single word than create debates regarding the questionnaire. Nevertheless, I report the reason why this item is left as an open-ended question and recommend this option to future researchers who work with children of that age.

7.2.6. Parallel forms

The participating students were assessed at two times in this study. The pre-test was administered at the beginning of the school year and the post-test at the end. The creativity activities included in the pre-test differ from these in the post-test. Despite having the same format, the activities used different verbal and the non-verbal stimuli. The equivalence of the pre-test to the post-test was examined via a pilot of the measurement tools. As I will explain in the next chapter, piloting of the assessment

tools suggested that the two verbal and the two non-verbal stimuli are equivalent to each other concerning their difficulty. Nevertheless, even if the two tests were not equivalent, having a comparison group ensured that any change in the performance of the intervention group is not due to the difficulty of the test items. Any change in the difficulty of the test affected both the control and intervention group.

Similarly, different items were used for the critical thinking assessment. The Cornell Critical Thinking Test Manual (Ennis, Millman & Tomko, 2005) recommended that the same test should be used both for pre and post-testing, because it is challenging, or not viable, to create a parallel test. However, I argue for the critical thinking assessment exactly what I argued for the creativity one. Given that there was a comparison group, even if the tests were not equivalent concerning their difficulty, this would affect the performance of both groups equally or in an unbiased way.

7.3. Creativity assessment

In the creativity test used in this research, there were two activities. One was stimulated by a verbal and the other by a non-verbal inducement. Both activities were open-ended. Guilford (1967) argued that divergent-production assessments need the examinees to produce their own answers instead of choosing from alternatives.

This first activity was scored based on three different elements: fluency, flexibility and prevalence. Then it was categorised in what is commonly used divergent thinking assessment (Plucker & Makel, 2010). In these assessments, a stimulus is provided to the individuals and they have to generate responses which are related to the provided stimulus. These assessments ask the individuals to generate as many responses as possible.

The first activity in the assessment of this thesis was a divergent thinking assessment created by Guilford (1956). He was the one who initially suggested a test where objects were given to students and unusual uses of these objects were sought by the assessment. He also created a similar assessment during which the students were asked to generate as many uses of the brick. This idea was later expanded by Getzels and Jackson (1962) who used a similar assessment called ‘Uses for Things’ also added other objects, pencils, paper clips, toothpicks and sheet of paper. During these activities, the students were expected to write as many uses for an object they can think of.

These authors also included in their assessment the phrase ‘Write down anything that comes to mind, no matter how strange it may seem’. This phrase was also used in this thesis, because it was judged that this phrase can liberate the students from the fear of writing something which might be a ‘crazy’ idea. Concerning the phrasing of this activity, Torrance (1988) recommended the use of the phrase ‘Try to think of something no one else will think of’ in order to increase the originality score. Originality is the same concept that this thesis calls inverse prevalence. Furthermore, Torrance (1988) explained that this phrase discourages students from cheating. However, this phrase was not adopted by this research because it is likely that it discourages them from also providing common responses. These were also needed since they were related to the fluency score, which was also graded.

To be more specific, this activity graded the fluency, the flexibility and the prevalence of the responses of the students. These are characteristics suggested and discussed in divergent thinking assessments (Davis, 1999; Getzels & Jackson, 1962). Fluency refers to how many uses of an object the students named without any further evaluation. The merit of the responses does not play any significant role at this level of judgement. Flexibility ‘refers to the number of different categories of ideas or the number of different approaches one takes’ (Davis, 1999, p. 215) when a problem or a stimulus is concerned. In other words, it involves a quality judgement of the responses an individual provided, and this judgment is within the responses of the same individual. Some individuals might have provided different uses which significantly differ from one another whilst other might have provided similar responses. Therefore, flexibility evaluates the quality of the responses in the individual. Finally, prevalence would have been called uniqueness by other researchers. This element refers to the responses of students when they are compared to the responses of the cohort. Some students might have provided more common responses whilst other might have thought something only a few or none of the other people have thought of.

There are two different indicators created and adopted for the scoring of prevalence by this research. This will be discussed in detail in the chapter which follows and focuses specifically on the grading of the activities. Guilford (1967) used the model of Structure of Intellect in order to categorise the test with the uses of a brick. He categorised this assessment in the divergent production and particularly production of units on a semantic level. This means that the students are called to produce separate ‘things’ with no links between their answers. Guilford (1967)

suggested that the flexibility measured by the task of the uses of brick should be called 'spontaneous flexibility' (p.143) because there is nothing in the phrasing of the activity to suggest to the examinees to take different approaches in their response. Thus, whoever does it, they demonstrate flexibility based on their own intuition.

The second activity of the creativity assessment included a non-verbal stimulus for the students, which was a half-complete image. The students were asked to complete the image and provide a title. This activity was based on the Torrance Test of Creative Thinking (Torrance, Ball & Safter, 2008). In the original test though, the students are required to complete many different activities, whilst in this assessment there was only one half-complete image. Furthermore, during Torrance Test the students are allowed to complete the images by combining more than one picture. Once more, this was not applicable in the assessment of this thesis since it required the completion of only one image. Furthermore, as it will later be explained the grading of the activity also differed from the Torrance guidelines. Therefore, it could be argued that even though this activity is based on Torrance Tests of Creative Thinking (Torrance, Ball & Safter, 2008) in fact it significantly differs from what the test would allow or measure.

The two indicators measured by this activity were the abstractness of the title written by the students and the resistance to premature closure. According to Torrance, Ball and Safter (2008) the first one is related to the individuals' ability to synthesise their thinking and in the highest level to capture the essence of information involved.

The resistance to premature closure is measured by the shapes that the students draw and whether these shapes were left open. This openness stands for the ability of the people to remain open and stand the ambiguity, which are personality characteristics of creative people as it was discussed earlier in Chapter 3. Even though the shapes are evaluated in that sense, the actual performance of the students in the drawing did not play any role. This assessment did not evaluate the performance of the students in the arts. As explained earlier, the assessment of this thesis is focused on evaluating creativity as a domain independent skill. Hence, evaluating creativity in the drawing is considered as evaluating creativity in the domain of arts and it is not an element that would be included in the general creativity.

Torrance Tests of Creative Thinking also evaluate the elaboration of the students. Even though this was not a separate element assessed by the assessment of this thesis, the adjusted scoring of the second activity enabled the rewarding of the

students when they elaborated their responses. For example, if the student drew a closed shape, that was scored with 0, whilst if the student added details in the enclosed shape, that drawing was scored with 1.

Finally, the relationship between these sub-scales used for the operationalisation of creativity should be discussed. Torrance removed the flexibility scale from his divergent thinking assessment, because it was too highly correlated to fluency scores (Plucker & Makel, 2010) and this is why there is no flexibility scale in the revised version of the Torrance Tests of Creative Thinking (Torrance, Ball, Safter, 2008). Also, researchers reported (Kim, 2006; Plucker & Makel, 2010) concern that the fluency score might contaminate the originality score (mentioned as prevalence in this thesis). It is expected to some extent to have some correlation between the two because they are elements of the same construct. Nevertheless, they should not be too highly correlated because they are still two different sub-scales of the construct. For this reason, the correlation between the sub-scales was examined in the results section of this thesis in order to decide to whether these scales measure the same or different things.

7.3.1. Literature review: Possible interpretations

In the last section on creativity, there are three basic claims identified in the relevant literature which are highly related to the current thesis. This thesis is not going to test these claims as hypotheses. Furthermore, these hypotheses are already based on data from other studies. However, these three suggestions might be possible interpretations for possible results.

7.3.1.1. The recent use of an object

Guilford (1967, p.327) summarised the results of experimental studies, particularly related to the psychology string problem, and he reported that recent use of objects in their common and conventional ways made it more difficult to think of unconventional uses of these objects. If this is the case, then the students of both control and intervention group will be affected when they try to think of uses for pencils. It is obvious that they use pencils more often than they do use bricks. Maybe this is the reason why Guilford (1956) introduced only the assessment with bricks, whilst the pencils were introduced by Getzels and Jackson (1962). If this is the case the performance of both groups will be affected and that could potentially explain more conventional responses as uses for pencils.

7.3.1.2. Age of the participants

Torrance (1962) reported that several studies including one of his own found that there is a decrease in the creativity of students at fourth grade and seventh grade. Particularly, fourth graders produced fewer stories, poems and inventions for the school's magazine compared to the other grades. He also explained that some of these students will lose their creative growth rather permanently and he discussed different explanations for this decrease in creative development, such as physiological ones.

However, social factors are probably the most crucial at this age, since stereotyping, competition and compromise are reported for the students of fourth graders. Similarly, with seventh graders, adolescence starts. This might cause insecurity and anxiety which does not facilitate creativity.

Guilford (1967, p.334) also reported that Torrance found the 'fourth-grade slump' and he reported that it takes two to three years for some students to recover, whilst some never recover. Guilford (1967) also discussed explanations for the phenomenon. He used the appearance of sex roles for the students of this age as an explanation of this. For example, independence is required for someone to be creative. However, independence appears to be a masculine characteristic and therefore girls are discouraged of being creative. Furthermore, the students are always presented with norms and therefore the pressure of conforming to the norms can restrict creativity.

7.3.1.3. Testing Conditions

Torrance (1988) examined carefully the testing conditions and the effect they have on performance of students when creativity assessments are concerned. He found out that students perform lower in overheated rooms with bad ventilation. Furthermore, particularly students examined in late May were found to perform lower compared to how they performed earlier in the year (Torrance, 1988). That applied even in the cases of children who received training. Therefore, this might affect the performance of the students in the post-test. Hopefully, this will not have a big impact on the schools in England because the weather is not particularly hot in June when the post-test was administered.

7.4. Designing the Critical thinking assessment

Having discussed the creativity assessments, it is important to explain the critical thinking items. My role in the construction of this assessment is more active compared

to the creativity assessment. Even though I based the ideas for the construction of the test on existing measurement tools, I created all the assessment.

7.4.1. Purpose

According to Cambridge Assessment (2017), the first step in constructing a great assessment is clearly stating the purpose of the assessment. An effective assessment fits the purpose which is designed for. The purpose of this assessment was to measure the critical thinking of cohorts of Year 5 pupils in order to evaluate the P4C intervention.

7.4.2. Why did I design a multiple-choice assessment?

The purpose of the assessment should determine its format of the assessment. For this research a multiple-choice assessment was used. Multiple-choice assessments have benefits that other types of assessments do not have. This does not mean that they are a panacea. Despite their limitations, tests of multiple-choice questions can be more reliable compared to other types of tests (Burton et al., 1991). First, multiple-choice questions are objectively scored. In other tests, like essays, there can be disagreement between the people marking the test (raters), which can increase the measurement error and lead to low inter-rater reliability. Moreover, a multiple-choice question does not take much time to be answered and, therefore, students can answer many multiple-choice questions in the same time that they could reply to a few open-ended questions or a single essay (Zimmaro, 2016). This enables the assessments to include more questions on the topic. Consequently, there is a broader coverage of the examined subject and therefore more representative results about the knowledge of the student (Burton et al., 1991). Furthermore, multiple-choice questions are not time-consuming to mark, and finally, they can focus on a specific topic.

Therefore, multiple-choice questions can make the assessment focus on critical thinking and reduce the construct irrelevance. Essay writing might enable access to the thinking process of students more than an assessment with multiple-choice questions. However, the students should be able to have a developed writing ability, which is not relevant to the construct of critical thinking. As Fisher and Scriven (1997, p.155) argued:

> we do not want to act as if critical thinking is the same thing as essay writing: there are many acute critical thinkers who do not write good essays.

Norris and Ennis (1989) also mentioned the benefits of using multiple-choice items in order to evaluate critical thinking. Even though there is no access to what students are thinking, the two authors recognised that one of the main advantages is 'the ease and the speed of acquiring reproducible scoring results' (p.28). Since the students do not need much time to answer a multiple-choice question, they can actually reply to more questions at the same time they would use to write an essay. Hence, the coverage of the domain can be broader, since different abilities and dispositions can be evaluated.

However, it should be recognised that during P4C sessions, the students are involved in a dialogue that does not have right or wrong answers. Thus, it could be argued that the dichotomous scoring of the measurement tools with right and wrong answers contradicts the nature of P4C itself. The scoring system implied that some answers are correct, and some answers are wrong. On the other hand, P4C introduces a way of thinking which any answer can be potentially right as long as a substantial justification is provided. Since I exposed the contradiction between the dichotomous scoring and P4C, it is apparent that open-ended questions being scored based on the extent of justification would have been more appropriate.

Despite the foregoing, I avoided using open-ended questions for two reasons. Firstly, it would be difficult to get an objective and reliable scoring system with different raters giving different grades for the same response. Secondly, even if the inter-rater reliability would have improved, that tool would include high construct irrelevance. It has been said that construct irrelevance is a source of invalidity and means the inclusion of irrelevant aspects from the measured construct (Messick, 1995, p.743). If the assessment included open-ended questions, several students might have struggled to express and justify their opinion in a written form. By asking the students to provide a written reply to express their opinion, the tool would also measure their writing ability. The exclusion of measuring the writing ability and hedging the construct irrelevance in the critical thinking test led to the construction of dichotomous multiple-choice items.

7.4.3. Why 3 alternatives in the multiple-choice questions?

The multiple-choice of the assessments have three alternatives (two distracters - or distracters - and one key answer) and this is not a random choice. A meta-analysis of multiple-choice questions reported that the research of the last 80-years demonstrates that two distracters are the ideal number of distracters for multiple-choice questions

(Rodriguez, 2005). The distracters should be equally plausible, and it is difficult to find equally more than three equally plausible answers. This means that it is better to use fewer distracters instead of using more distracters with some of them being not functional. Tarrant, Ware and Mohammed (2009) recognised that teachers believe that fewer distracters might lead to easier assessments due to the increased likelihood for guessing. For this reason, they analysed tests and they demonstrated that by removing all the non-functional distracters, the performance of the students increased by only 1%.

Moreover, there is an additional reason for using three alternatives. Specifically for the inference question, Norris and Ennis (1989) suggested that in the Cornell test three options (true, false or neither) instead of five (true, probably true, false, probably false or neither) were chosen. This was because choosing the right option amongst five alternatives might have been too confusing. These five options are used in the Watson Glaser test. According to Norris and Ennis (1989) this slight differentiation offered by five options might lead some students to choose the wrong answer, while they would pick the right one if they had only three due to mostly background beliefs. Thus, due to this type of problem three options were chosen and for consistency reasons that would have been followed in all the assessment. Consequently, three options were chosen for the multiple-choice questions of the assessment.

7.4.4. Guidelines for constructing good multiple-choice items

The discussion of the choice of alternatives in the multiple-choice questions is only an example which can demonstrate how carefully the questions were constructed. Norris and Ennis (1989) recommended that the multiple-choice questions in critical thinking assessments should be carefully designed in order to have only one correct answer provided among the alternatives. Downing (2006) claimed that good multiple-choice writers are properly trained and not born.

For this reason, I have read carefully how effective multiple-choice items are designed. In my role as an education consultant, I wrote *'Multiple-Choice Items: A guide for teachers'*. In this document, I summarised guidelines about effective writing of multiple-choice questions (Ventista, 2017) based on existing evidence. When applicable I followed these guidelines for this assessment. I do not aim in this section to explain all the guidelines that I followed. However, some examples are mentioned in order to clarify the process that the items were constructed.

- The stem of the items did not contain irrelevant material (Brame, 2013). There are many reasons for this. For example, making the stem longer to read increases the reading ability which is demanded by the students and therefore increases the construct irrelevance.
- No tricky and opinion-based questions (Haladyna et al., 2002; Zimmaro, 2010) were included. The purpose of multiple-choice questions was to assess critical thinking and not to trick the students. For the same reason, negative phrasing was avoided in case it was confusing.
- Alternatives were mutually exclusive (Brame, 2013) and not overlap (Zimmaro, 2010). As Norris and Ennis (1989) recommended there was only one correct response for each question.
- One of the common mistakes in multiple-choice questions is that greater detail is offered for the correct option (Tarrant & Ware, 2008). All of the alternatives were similar in length and they did not provide clues about the correct answer.

Every decision in the construction of the multiple-choice questions was justified. Even the names of the characters included in the thinking problems were chosen carefully. There was an attempt to also include names which were used by various cultures and not only British culture, because the latter might have introduced cultural biases in the assessment.

7.4.5. Challenges in designing the assessment

Except the writing of multiple-choice questions, there were other challenges faced when designing the critical thinking assessment. Additionally, the second enormous challenge was how to motivate the students to read these texts and think to solve the problems when they know that this type of testing does not have consequences on their grades. To engage the students, I inspired the item format by a test trialled for this age group - the Cornell Critical Thinking Test Level X (Ennis & Millman, 2005). This test narrates a story in which a group of people arrives on a newly discovered planet in a year in the future. However, they ended up losing contact with people from Earth. For this reason, a second group is sent from the Earth to find out what happened. The examinee is presented as a member of the second group. While the story develops, different test questions are asked to the students. The end of the story is provided in a few lines at the end of the test.

I argue that this type of testing can be extremely engaging for the students grounded on the feedback I received when I trialled the test of appraising observations (Norris & King, 1984) with secondary school students (Ventista, 2018a; Ventista & Coe, 2015). Specifically, the test of appraising observations has a similar setting. This time, the examinee is presented in the first section as a member of a police team which investigates an accident and in the second section a member of a group of explorers. While I administered this test, many of the students told me that they found the test engaging, and it made them feel like being detectives. Thus, creating a story and interesting context for the problem can be engaging for the students.

However, I supposed that it is possible that the story can be evenly disengaging. I received the previous positive comments mainly from boys. This can be due to the fact that the one of the two schools I administered the test had only boy students. However, in the other school I did not find that the girls were equally engaged with this test. My sample had only 15 girls, and they seemed to be more engaged with the creativity tests. Since the sample in that research was small, I cannot reach any generalisable conclusion supporting that girls are not equally involved. Examining bias in the performance with such a small sample of girls would not lead to trustworthy conclusions.

But what if the girls are not interested in space and adventure in a new planet? Someone could argue against my thought by advocating that such a perception is sexist. Is it biased to infer that the girls are not equally fascinated by these adventures? I recognise that there is this possibility. Although I disagree, I also understand that our society is not still free from biases about the interests of the boys and the girls. Therefore, as an alternative, I decided in this critical thinking test to include problems with various contexts. I did not want to assess the students restricting the problems in solely one context. I used contexts which can be considered gender-neutral to hedge the presence of potential gender bias. Topics such as family, music, holiday or chocolate cakes, can be of interest of both genders.

This careful consideration of the factor of context in the thinking problems did not take place only due to the attempt to engage students with the problems. Context was judged important for critical thinking problems. First, in his definition of critical thinking, Lipman (2003) suggested that critical thinking is sensitive to context. In order for this to be assessed, the problems presented have to be placed in a context. Furthermore, critical thinking is not perceived only with the idea of the strict formal

logic. As Fisher and Scriven (1997) argued one of the flaws of formal logic was that the presented vacuous examples, while in fact the examples are necessary for critical thinking. This idea implies that the context is necessary for critical thinking. Consequently, the emphasis on the consideration of context is not only based on its role for the students' engagement. The context has a functional and necessary role for critical thinking problems. In the next sections, I will discuss the construction of each item separately.

Table 7.2. Simplified Version of a Test Blueprint for Critical Thinking Assessment

Critical Thinking Elements	**Pre-test**	**Post-test**
Inference	Problem 1	Problem 1
Evaluation of the Argument and Credibility of Sources	Problem 2	Problem 2
Deduction	Problems 3 and 4	Problems 3 and 4
Assumption Identification	Problem 5	Problem 6
Problem-Solving	Problems 6 and 7	Problems 5 and 7

7.5. Content of Critical Thinking Assessments

7.5.1. Inference

The first thinking problem examined the inference. In the item the students were asked to draw a conclusion which was warranted with the evidence given. In both pre-test and post-test, the students should have replied that they do not have adequate evidence to reach a conclusion. For example, the ticket from a music concert does not entail that a person takes guitar lesson. It might suggest that the person is interested in music, because otherwise the person would not like to take the lessons. However, it does not lead to a warrant conclusion.

The inference is commonly used in other critical thinking tests. For example, it is the first component of the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 2002), which is a test for graduates. Cornell test Level X (Ennis & Millman, 2005) also starts with a similar type of activity. The activity I constructed resembled more the Cornell's test, because it included three alternatives instead of five. For this age group, I judged that it would be unnecessary to offer so many options.

Sample Item 7.1. Inference Item used in the post-test.

PROBLEM 1: DOES YOUR BROTHER LEARN THE GUITAR?

Today your mother says 'I think your brother is having secret guitar lessons. I found a ticket from a music concert when I cleaned his room'.
When you hear this, you:
A. think that your brother is having guitar lessons.
B. think that your brother is not having guitar lessons.
C. cannot decide if he is having guitar lessons or not.

7.5.2. Evaluation of an argument and credibility of sources

The evaluation of argument in this assessment included a judgment about the credibility of an argument. Robert Ennis used in his tests activities which required judgement on the credibility of sources. The credibility of sources is also the second part in the Cornell Critical Thinking Test Level X (Ennis & Millman, 2005), while statements from different people are presented and the student should judge whether the advice is credible. The specific item I constructed was based on the test of appraising observations (Norris & King, 1984) and the Cornell Critical Thinking Test Level X (Ennis & Millman, 2005).

In the case of this assessment, the statement of an authority was presented. The statement of an authority might be the most trustworthy statement. However, it could also be the case that a statement of an authority is used in an erroneous and misleading way. One of the most common fallacies is what is called Argumentum ad Verecundiam, (Harrison-Barbet, 2001, p.51) when the authority's opinion is used to suggest a conclusion on an irrelevant topic. A critical thinker should be able to judge whether the authority is the relevant authority and able to offer a credible statement.

Sample Item 7.2. Evaluation of Sources used in the post-test.

PROBLEM 2: WHO DO YOU BELIEVE?

Sarah has many headaches and she decides to visit a doctor. The doctor asks a series of questions, and tells her: 'You should drink more water'. When she comes out of the doctors, she meets a friend. She explains that she has just been to the doctor for the headaches. Her friend says 'Every time you are thirsty, you should drink green tea - not water. People say that green tea helps to reduce headaches'. Whose advice should Sarah follow?
A. The doctor's
B. Her friend's
C. Both the doctor's and her friend's

This problem could be split into two different items, each of which would ask the students to evaluate the arguments. For example in this example the doctor could not be an authority to suggest on the person to which internship should accept for the summer! In other words, the authority in this example is used without a fallacy. A critical thinker should be able to judge that the doctor's advice should be trusted in this case because the authority in this case is the relevant authority to be trusted for the health advice.

Then, there is also the advice offered by a friend. The friend suggests drinking green tea. The friend suggests this, because (s)he has merely heard it from other people saying. The knowledge, which is shared in a community, tends to be trusted by people. However, the evidence that green tea helps in this example is not sufficient. There are many examples of society knowledge which can have a detrimental impact on health or other aspects of life when it is trusted. For example, in Vienna in the mid-eighteenth century, it was accepted that the ideal food for infants was not breastfeeding, but bread boiled with water or beer with added sugar. With almost 60% of babies dying – possibly due to dreadful feeding - Mozart himself lost four of his six children (Jenkins, 1995). A critical thinker nowadays should be able to judge the credibility of the source of knowledge and should search for additional evidence before trusting a health recommendation by the society.

For the item about the judgment of credibility of source, comparative judgment items were used. According to Norris and Ennis (1989) the credibility items are

categorised in comparative and non-comparative judgment. The first type offers two statements from two different people and requires by the students to judge which one is more believable or judge that they are equally believable. The first type offers only one statement and the students should decide among the options of trusting the statement, not trust or being indecisive about it. Both in the pre-test and the post-test, comparative judgments were used. However, there was another technique used in this test. The two statements provided are contradictory. Thus, they cannot be both correct and possible to follow them. This is compatible to the recommendation about the construction of good multiple-choice assessments, which states that the alternative in Multiple-Choice questions should be mutually exclusive (Zimmaro, 2010). Therefore, students cannot follow both pieces of advice at the same time. With common sense critical thinking, students should be able to judge that they cannot act in both ways.

There is one main reason I provided contradictory statements. It has to be recognised that the problem with the credibility items are their dependence on the background beliefs of the examinees about how physical, social and cultural worlds operate (Norris & Ennis, 1989). Thus, I judged that at least one of the options (the C) in both cases could be excluded whether the student could judge that they cannot follow two contradictory pieces of advice. This requires judgment of sources without introducing any background beliefs.

Nevertheless, the problem of background beliefs might still remain. Norris and Ennis (1989) suggested that a follow-up question might reveal the thinking process of choosing an answer and therefore the rater might be able to explore potential background beliefs which might have led to the wrong answer instead of lacking critical thinking ability. This option was not followed because it was not judged appropriate. On the one hand if an open-ended follow-up question was used, it would have introduced disadvantages for students with difficulty in the writing ability. On the other hand, if the follow-up question was a multiple-choice question, the alternatives might hint the right answer on the first question and some students might have returned to change their initial response.

7.5.3. Deduction

Reasoning was one of the parts of working definition of critical thinking for this thesis. Deduction was chosen as a type of reasoning which can be easily assessed in a multiple-choice format. Ennis has also included deductive reasoning in the Cornell

Test for Critical thinking (Ennis & Millman, 2005). Furthermore, critical thinking assessments sometimes include spatial reasoning, such as CAT assessment. However, this reasoning was not included in the assessments of this thesis. It is adopted the approach of Fisher and Scriven (1997, p.51) who supported that critical thinking requires the ability to deal with a language. This means that language is crucial for critical thinking to be evaluated and therefore the problems in this assessment are linguistic.

There are different types of deductive syllogisms. For these reasons, I decided to include one simple and one more complicated syllogism. My initial thought was to incorporate a syllogism with what is usually called universal premise (Evans, 2005). The sentences starting with 'all' and 'no' are introduced in the first chapter of *Harry Stottlemeier's Discovery* (Splitter, 1992). I started thinking an example with cats with the universal premise 'all cats have four legs' and ask the students to think that deductively that as all cats have four legs, then the cats that the students see have four legs too. That was the first item I constructed:

> In your grandmother's yard you see some cats.
> All cats have four legs.
> These are cats.
> Therefore,

A. the cats you see in your grandmother's yard have four legs.
B. the cats you see in your grandmother's yard don't have four legs.
C. if they have four legs, then they are cats.

Nevertheless, I started thinking that this item should not be included because all the students would answer this correctly. I believed that all the students would provide with the correct answer not because they are capable of deductive thinking, but because they have encountered four-leg cats in real life. In a similar paradigm, Evans (2005) supported that people can reach a conclusion merely because they have experience on this topic. In the example by asking the students to conclude that all the cats have four legs, I would not measure deductive reasoning, but their knowledge grounded on their experience. As this fact is known to all the students, all the students will answer this item correctly and the question would not have added information on examinees' critical thinking ability.

Furthermore, if an item is answered correctly by all the students who take the test, then it is not discriminative. Discriminative is not an item which is biased against a particular group, but an item which reveals an existing difference between the people who take a test (Koretz, 2006, p. 27). Koretz (2006, p.28-29) suggested that when relative proficiency is examined and differentiation between performances is sought, then a test should include discriminating items. These items are not too easy or too difficult.

What is more, the aim was to evaluate deductive reasoning by excluding the experience of the students. This could be achieved by asking them to deduct with unreal premises. That was the next item I constructed;

> In your grandmother's yard you see some cats. All cats have three legs.
> Therefore

A. the cats you see in your grandmother's yard have three legs.
B. the cats you see in your grandmother's yard don't have three legs.
C. if they have three legs, then they are cats.

This is an outstanding example of the problem that I was referring at the theoretical chapter of this thesis. In deduction, false premises can draw a valid conclusion. The correct answer in this problem is A. However, it is likely that many – and possibly the majority of the – students would have chosen option B. This is not because the students cannot deduct, but because as Evans (2005) claimed, people tend to reject a valid conclusion when this is not in accordance with their experience. This is not an indicator of lack of deductive ability and definitely not a lack of critical thinking. On the contrary, I assume that the students who might choose B would are more critical than the students who choose A in the example with the three-leg cats. My previous findings (Ventista, 2018a) suggested that context and stimuli closely related to students' everyday life could restrict the critical thinking that demonstrate in the assessment instead of facilitating it.

Consequently, I rejected the idea of including an argument with a universal premise. After rejecting the inclusion of a syllogism with a universal premise, I decided to include a syllogism known in philosophy as modus tollens.

Given: p→q (if p, then q)

Given: ¬*q* (not q)

Therefore, ¬ p (not p)

Sample Item 7.3. Deduction Item (Modus Tollens) used in the pre-test.

PROBLEM 3: THE MEETING

Every time I meet Robert, we go to the cinema to see a film. I did not watch a film yesterday. This means that

A. I met Robert yesterday.
B. I did not meet Robert yesterday.
C. I might have met Robert yesterday.

At the end, in both pre-test and post-test I included two reasoning items. The one was a modus tollens and the other a more complicated deductive syllogism. Both of the items belong to the category comparative-judgment approach, as defined by Norris and Ennis (1989). In the comparative approach, the students are given conclusions and they are asked to choose the best one among them. The non-comparative approach involves providing just one possible conclusion and asking the students to judge whether they think is true, false or they cannot decide. I chose the comparative-judgment approach because I thought that this usually resembles the situations anyone faces in real-life. There are usually many alternatives from which somebody is called to choose a conclusion.

7.5.4. Assumption Identification

Assumption identification is a typical type of questions which can be found in critical thinking assessments, such as the Watson and Glaser (2002). In this test the authors instead of asking the examinees to identify assumptions, they asked them to decide on whether a statement is assumed or not. In the assessment of this thesis, the students were asked to identify assumptions. For example, in the post-test they are asked to identify the assumption in the argument of the head teacher.

Sample Item 7.4. Assumption Identification Item used in the post-test.

PROBLEM 6: AN ANNOUNCEMENT

Today the Head teacher said: 'Every afternoon there is heavy traffic in front of our school, and a student might be hit by a car. To make sure that no student will be hit by a car, please ask your parents to avoid driving on the road in front of the school entrance in the afternoon' Students reacted differently to this message. Which comment makes more sense?

A. 'The cars are not driven only by our parents. Other people drive on this road, too'.
B. 'The drivers are always careful, so it is unlikely that a student will be hit by a car'.
C. 'The road in front of our school should only be busy in the morning. Not many students are walking in the morning'.

There is one typical principle about test development which was violated in this item. Even though there should be one correct answer and only the students who lack the knowledge should be confused by the wrong possible choices provided in the multiple-choice item (Haladyna, 1994, p. 80-81), this was not the case in this assessment. It did not have specific knowledge that the students should have demonstrated and for this reason the correct answer did not have to be apparently declared. The correct answer is the better option in this item. Even though it is much more likely that the cars in front of the school are driven by the parents of the students, there is still a possibility for the cars to be driven by other people who might return from their work and they work near school. Statement C is logically erroneous since in the announcement, there is no reference about what happens in the morning.

The options provided in this multiple-choice item were carefully designed. When choices are given to the students to choose the best, it has to be reassured that the correct answer is not too discernible compared to the wrong answers. It is also not advisable to include an answer which could be too misleading for the students. It has been advised that the test constructor should provide possible answers in the multiple-choice items but should not aim to trick the students (Haladyna, 1994). This could be

the case for a knowledge test, but it is more challenging to keep the balance with a critical thinking test, because as no prior knowledge is demanded the answer should not be too obvious. Therefore, for the measurement tool of this research a distractor, which is any wrong option given in a multiple-choice test, should not be evenly conceivable as the correct answer. Distractors should be equally believable (Brame, 2013; Haladyna et al., 2002). In other words, in piloting I would conceive as an unsuccessful distractor any option that confuses more than half of the students taking the test.

For the assumption identification item, the guideline provided by Norris and Ennis (1989) were followed. These authors recommended that a conclusion which could be drawn by the given statements should not be included in the alternatives. By excluding a potential conclusion, the students who did not understand adequately what the question requires are not penalised. In this sense, only potential assumptions were included in the alternatives.

7.5.5. Problem-solving

Two items to include students' problem-solving ability were included. This is compatible with the ideas of Sternberg (1986) who included problem-solving in the critical thinking. Furthermore, a subsection about problem-solving can be found in popular critical thinking assessments, such as Halpern Critical Thinking Test (2010) and New Jersey Test of Reasoning Skills (Shipman, 1983).

However, Pritchard (1992) criticised the association of critical thinking with problem solving. He argued that it is problematic, because critical thinking problems necessitate a sole and 'the most logical choice' (p.92) by rejecting more productive, smart or creative replies. To resolve this issue, I tried to pursue using problems that, even though they demand the students to choose amongst three logically correct answers, one answer is clearly the best one and the students who think more productively will not provide a different one and be underscored.

I argue that a problem-solving item could evaluate either the creativity or the critical thinking of the students. In the particular items I constructed, the different alternatives are provided to the students. Therefore, they have to use their judgment and evaluate which is the best solution for the problem. In this sense, the problem has been turned into a judgment problem and therefore a critical thinking problem. In case the same problems were open-ended, the students would have been required to

generate the solutions and it would be acceptable for the same problems to be used to assess creativity.

7.6. Psychometric properties

Having decided on an assessment for creativity and having constructed the measurement tool for critical thinking, psychometric properties of the assessments are examined. Delphi report for critical thinking (American Philosophical Association, 1990) recommended that any critical thinking assessment should be examined for content validity, construct validity, reliability and fairness. The first properties are examined in this section, while fairness and biases can only be examined after the collection of data.

7.6.1. Reliability

According to the American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014), reliability is 'the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker'. Therefore, reliability can be considered as replicability. According to Koretz (2006), reliability is also consistency of measurement. Consequently, reliability can be consistency across different occasion, but also consistency within the test. The latter would mean that the test measures the same construct.

Norris and Ennis (1989) discussed the reliability as a quality indicator of a critical thinking assessment. Particularly, they explained that reliability as consistency from one occasion to another is a desirable element. Nevertheless, it is not possible to brainwash the students and examine this element effectively. Therefore, different methods are used, such as split-half reliability and Kuder-Richardson reliability. However, measuring reliability with methods such as the Kuder-Richardson reliability are not always appropriate of critical thinking assessments. They wrote (p.46):

> Kuder-Richardson reliability estimates are high to the extent that scores on individual items on the test correlate with one another, and low to the extent that they do not correlate. This sort of estimate may be quite inappropriate for tests of various aspects of critical thinking because there is not theoretical reason for believing that all the items on such tests should correlate highly with one another.

According to the same authors, even reliability of 0.65 could be considered adequate for critical thinking tests (p.47) because almost each item examined a different aspect of critical thinking. Hence, this test was not checked for reliability as internal consistency. Internal consistency is also examined by Cronbach Alpha in SPSS (Wilson, 2005). If the test is reliable then it should be repeatable (Koretz, 2006). A distribution between the scores each time a student takes a test is expected, but if the test can be done many times it is expected gradually to approach the correct score without measurement error (Koretz, 2006). However, as it cannot be a "brainwashing" the Alpha splits the test in two halves, assumes that the two halves are two different repetitions of the implementation of the test and examines if there is consistency (Wilson, 2005). However, in a measurement of a multi-facet construct it would not be useful to examine the consistency between the different problems. High performance, for example, in deduction does not necessarily entail high performance in interference item.

Similarly, in the creativity assessment, good performance in one of the aspects of creativity does not necessarily imply high performance in a different aspect. As a result, reliability as internal consistency was not of an interest for the quality of creativity assessment.

Concerning inter-rater reliability though, it was carefully examined particularly in the case of creativity activities. Multiple-choice items are marked in a reliable way, since they have a single correct answer, and this is also one of their main advantages. However, creativity assessments involved a level of judgment and therefore there were many different ways of ensuring the inter-rater reliability.

7.6.2. Validity

Concerning validity, Norris and Ennis (1989) recommended three types of validity that should be examined in the case of critical thinking assessments. These are the criterion-related validity, the content-related validity and the construct-related validity. I am going to discuss each of the types separately.

First of all, concerning the criterion-related validity, Norris and Ennis (1989) recommended the correlation of test scores between the discussed test and a different critical thinking test. In this sense, a test is valid when the students who perform highly in the test, they will also perform highly in another critical thinking assessment. Vice

versa, if students perform poorly in the critical thinking test, then they will also perform poorly in another one. In this sense, the assessment is valid since it measures what is supposed to measure as confirmed with the correlation of the test scores.

This type of validation was not possible for the critical thinking assessment used for this research. A critical thinking assessment was designed because of the lack of existing critical thinking assessments for Year 5 students. Therefore, examination about the criterion validity did not take place, because there was lack of appropriate critical thinking assessment for this age group. Furthermore, even if they were available tests, they would have to measure the same exact aspects of critical thinking. As I have found with multi-trait and multi-method matrix in a previous research (Ventista, 2018a), when I correlated test scored obtained in critical thinking assessments they were not correlated highly when they measured different aspects of critical thinking.

The second type of validity discussed is the content-related validity. This type of validity is based on experts' judgement. In order to reassure, that the tests were valid when their content was concerned, I based the working definitions used for the assessments on a solid theoretical background.

Finally, concerning the construct-related activity, it refers to the construct, which is defined as the 'underlying abilities, dispositions or traits of human beings, as opposed to directly observable characteristics' (Norris & Ennis, 1989, p.50). These underlying abilities are not directly observable. Particularly, with the use of multiple-choice items the access to the thinking process is even more limited (Norris & Ennis, 1989). However, these constructs have been carefully operationalised based on the working definitions. I created the working definitions for both constructs on extensive bibliography of definitions of the two skills as it became apparent in the definitions chapter. The working definitions have strong theoretical background and they expressed clearly to facilitate the measurement of the skills. For each aspect of the working definition, there is an activity or part of an activity which represents them in the final assessments used for the purposes of this research.

Concerning the creativity assessments, I have correlated the scores from the first activity with a creativity activity of partially completed images and I found the creativity activity 1 valid by using a multi-trait multi-method matrix (Ventista, 2018a). This means that activity one has convergent validity with other creativity assessments. Furthermore, this assessment had discriminant validity when it was compared to

critical thinking assessments. Therefore, the activity one could be considered a valid assessment of creativity based on the method of validation of multi-trait and multi-method matrix (Campbell & Fiske, 1959).

Regarding the second activity it was based on the Torrance Test of Creative Thinking (Torrance, Ball & Safter, 2008). Consequently, the psychometric properties of the original test Torrance Test did not apply. However, this activity corresponds to the working definition provided in the second chapter and in this sense the test could be considered valid since it measures what it claims to measure.

Messick (1995) criticised the approach of validity which treats these three parts of validity as separate elements and he unified validity. According to his approach, a unified concept of validity considers the test interpretation and use. This refers to the consequential aspects of the concept of validity.

The specific assessments were used to evaluate the effectiveness of P4C intervention and there was no judgement made for any individual. Therefore, it did not have any consequences for individual students. If there were serious consequences for individuals or the programme, then a panel of experts would be required in order to ensure the validity. In this case, the extensive bibliography research, the operationalisation of constructs and the piloting were sufficient for the use of the assessment for the intended purpose.

Concerning validity, Kane (1990) also focused more on the inferences drawn by the test scores and supports an approach where validation takes place as an argument-based approach. He did not focus on test scores but how these tests scores are used. Test scores are used in order to make decisions or draw conclusions about the performance of individuals or groups and sometimes in order to predict their future performance. In that sense, test scores use inferences to lead to conclusions and as Kane (1990, p.6) argued ‘The reasoning from a score to one or more such conclusions is necessarily based on assumptions’. These inferences and assumptions used to lead from a test score to a conclusion establish an argument.

According to his argument-based approach, Kane (1992) recommended four stages in the validation process: a) recognition of the decision or decisions to be based on the particular assessment scores, b) specification of the inferences and assumptions which lead from the test scores to these decisions c) identification of competing interpretations and d) identify the evidence that support the interpretations and assumptions and reject the competing interpretations. Finally, concerning the

validation as discussed by Kane, this process took place when these assessments were used. The assumptions made about the performance of the students are always stated clearly and there was a careful examination for alternative interpretations and construct irrelevance. When the results of the assessments are interpreted alternatives were provided. Furthermore, alternative interpretations and how these interpretations were faced are mentioned in the limitations section. The assessments were used as indications for whether P4C is effective. According to this, the assessments are valid, because they can sufficiently provide indicators about the effectiveness of the programme.

7.6.2.1. Construct irrelevance variance: Reading ability

Critical thinking tests usually require the reading of long texts to reply to their answers. However, the test should not include construct irrelevance (Messick, 1995) and the ability of reading should not affect the test performance (Hewitt, & Homan, 2003). The immense challenge with this test is how to measure thinking skills without measuring reading ability. For this reason, I tried to avoid the incorporation of extensive texts. The students were still asked to read passages whose content was improved by following the guidelines for improving the functioning content of items (Gronlund, 1982, p. 30); avoid including complex sentence structure, wordy and vague vocabulary and statements and race/ethnic/sex-biased content. After deciding on the construct and the items, I calculated the reading difficulty for the items using an online software (https://readability-score.com/). The software gives the Flench Reading Ease score. I adjusted the reading difficulty of the items to make them suitable for Year 5 students. Therefore, each item had appropriate Flench score for students of this age group (Table 7.3).

Table 7.3. Flench Reading Ease Score for the Readability of Items in Pre-test and Post-test.

	Type of Item	Item	Pre-test	Post-test
	Creativity Activity	Activity 1	84.2	87.1
		Activity 2	78.5	78.5
	Thinking Problem	1	93.8	83.2
		2	86.1	91.3
		3	90.7	102.4
		4	85.5	89.1
		5	88.8	87.8

		6	89	90.4
		7	87	100.1
Average Readability Score			87	90

From the table related to the readability scores it can be found that the post-test is slightly easier to be read compared to the pre-test. This is not a big difference and the two tests were judged equivalent. However, the readability might play a slight role because the pre-test is slightly more difficult than the post-test. This information can be combined with the fact that the students were younger at the beginning of Year 5 when they had to complete the pre-test, whilst they were about to complete the Year 5 when they completed the post-test. Therefore, the difficulty in readability combined with the younger age of the students might have led to a poorer performance of the students in the pre-test. That limitation was examined and can be found in the results section.

Despite the fact that excluding construct irrelevance variance entails that the reading ability or the Maths ability or any other ability should not be included in the final score of critical thinking tests, there were a few assumptions. First, even though high level of reading ability is not required for the completion of the assessments, I assume that the students were able to read the instructions and the problems and comprehend them in a sufficient level. Secondly, even though the mathematics ability was not included, there was a simple mathematic understanding required for the tests to be completed. For example, there was problem which required from the students to add hours to conclude whether it was day or night. Finally, there was the assumption that students have basic cultural background knowledge. For example, to reply a thinking problem they should have known what a guitar is and have the cultural understanding that somebody can attend a music concert without knowing how to play a musical instrument or similarly for the creativity task they should have known what a brick is in order to suggest different uses.

7.7. Criteria for Evaluating Tests

So far, all the decisions taken for the construction of the measurement tools have been presented. Furthermore, the psychometric properties of the test have been explored. However, before concluding this chapter, I will demonstrate how the measurement tools were constructed satisfy the main criteria used for evaluation of the quality of the

tests suggested by Fisher and Scriven (1997). Particularly, Fisher and Scriven (1997) suggested seven evaluation domains:

- Construction
- Administration
- Suitability
- Coverage
- Scoring Process
- Interpretation
- Report

Some information about these areas has already been discussed. However, it is important discuss them further based on the recommendations of these authors.

7.7.1. Construction

The construction of the assessments was not a product of teachers' effort. There were no additional supplies required, since this assessment was a common paper and pencil test. Therefore, the paper assessments were posted to the students and it was expected that the students would have a pencil to complete them. Fisher and Scriven (1997) also referred to environmental considerations. They understood that the multiple-choice assessments require more paper compared to other type of assessments and their use should be justified. I have already explained how multiple-choice questions reduced construct irrelevance and therefore they were the appropriate format for the assessments of this thesis. Furthermore, the assessments were not lengthy. Thus, environment was considered, but the use of multiple-choice questions was crucial.

7.7.2. Administration

For the administration of the test, there was a standardised process to be followed. A clear guide was sent to the schools with the forms. In the guide there was a clarification that 'During the administration, you can help your students to understand the questions (e.g. you can read the questions to a student with low reading ability or visual impairments) but you should not provide them with hints to respond to the questions'. The full administration guide, post-administration form, the assessments and the letter sent to the schools can be found in the appendixes 2.a. - 2.e.

The assessments did not have provision for students with severe special education needs and as a result special education needs schools were not recruited in the intervention and the comparison group. However, it became clear to the schools that they could offer the same assistance offered to the special needs students in other assessments in order to complete this assessment. It has already been discussed that the duration of the test was age appropriate and the stress levels were not judged high since there were no consequences for the students and the assessments were anonymous.

7.7.3. Suitability

The assessments were evaluated based on the language difficulty and their appropriateness for students of that age. Then, as it will be discussed in a later chapter they were piloted in order to confirm that the content of the assessments was suitable for the participants. Fisher and Scriven (1997, p.45) argued that in order to consider critical thinking there are three elements that should be taken into consideration: what the person in question can manage, the background and particularly the education and their age. Critical thinking demonstrated by a young child is not the same with the critical thinking of the adult, because adults might demonstrate some skills as a result of educational background and not necessarily because of critical thinking. Hence, as a concluding thought, it has to be expressed that the content of the critical thinking problems was constructed in a way which considered the critical thinking abilities that students of this age could demonstrate. However, this issue was treated seriously, and it is one of the main reasons that the assessments were piloted.

7.7.4. Coverage

Having discussed the working definitions and each item of the assessment, the coverage of the tests became clear. It was discussed that the test indirectly covers a few more areas and it is assumed that the students were aware of these domains, such as basic reading and writing ability and basic knowledge of the context of the problems.

7.7.5. Scoring Process

The methodology of the scoring process for creativity is described in great detail in the next chapter. The scoring of the multiple-choice assessments appears to be easier, while the scoring of the creativity was more challenging. I scored mainly the assessments, but during the process I used various raters to increase the interrater reliability particularly for the marking of the creativity activities.

All the items in the critical thinking test were scored with 1 if the item was answered correctly and with 0 if the answered provided was erroneous. Therefore, all the items in this test were dichotomous (Wilson, 2005, p.86). One answer for each item was correct, and it was discussed in the previous chapter how this scoring system can contradict the idea of pluralism in a community of enquiry.

Reasoning and problem-solving questions had two questions in the assessment. There was no intention for a double weighting for the calculation of critical thinking overall because all skills were judged equally important. Hence, the calculation of critical thinking for pre-test avoided the unequal weighting.

$$Critical\ Thinking_{Overall} = \frac{CT1 + CT2 + \frac{1}{2}\,CT3 + \frac{1}{2}\,CT4 + CT5 + \frac{1}{2}\,CT6 + \frac{1}{2}\,CT7}{5}$$

Similarly, for the calculation of the critical thinking in the post-test

$$Critical\ Thinking_{Overall} = \frac{CT1 + CT2 + \frac{1}{2}\,CT3 + \frac{1}{2}\,CT4 + \frac{1}{2}CT5 + CT6 + \frac{1}{2}CT7}{5}$$

7.7.6. Interpretation

The interpretation is the domain that Fisher and Scriven (1997) used in order to refer to the interferences drawn from the test results. These assessments are used for interferences only about the Philosophy for Children programme effectiveness and there is no interference for the performance of an individual or a school.

7.7.7. Report

It would be unethical to require time and effort from the schools and not report them the results of this study. The reports are compatible with the interpretation of the assessments by this research. Since the interpretation of the assessments were used in order to evaluate Philosophy for Children, the reports sent to the schools did not refer to individuals, but only to the results of the study and explained how Philosophy for Children schools were compared to the comparison group schools.

7.8. Chapter Summary

This chapter explained why I decided on the implementation of one 30-minutes test which assesses firstly creativity and then critical thinking. To be more precise, this

assessment demands a reasonable amount of time and easy administration. Nevertheless, the significance of excluding the possibility of students performing poorly to irrelevant constructs, such as reading ability or spending more assessment time on the other construct was discussed. I decided to provide concrete guidelines and standardise splitting the time within the two constructs evaluated in the same test.

This chapter did not discuss about the scoring guidelines of the assessments and this is something to be discussed in the following chapter. This chapter, however, discussed how the assessments were appropriate for the age of the participants. The assessment time took into consideration the age of the students and their reading ability. However, something which should be mentioned - if this has not been made clear already - is that the content of the problems were also appropriate for the students taking the test. As it has already been discussed, this is based on recommendations of researchers, such as the Think aloud technique which gives access to the thinking of the students when they reply to a multiple-choice item (Fisher & Scriven, 1997) or interview students similar to the targeted examinees to identify their background beliefs (Norris & Ennis, 1989). Therefore, the measurement tools cannot be considered complete until they are piloted in a population similar to the targeted participants. In a following chapter, this piloting is discussed.

8. Methods: Grading System for the Assessments

This chapter presents the grading system implemented for the measurement tools used for the second and the third research questions of this thesis. This chapter presents the scoring process for the creativity questions in detail. The grading system for the creativity activities was developed specifically for this thesis. There is a detailed explanation of the grading system because there is no source to be referenced. The critical thinking assessment included only multiple-choice items and as a result the grading was straightforward process. Multiple-choice items can be marked objectively since a student responded correctly or not.

Two of the main aims of the marking were to avoid bias and ensure consistency. On the one hand, multiple-choice questions were scored as right or wrong. On the other hand, the creativity assessment had open-ended questions. As a researcher I had to mark these assessments. In order to avoid marking bias I scored these assessments with blind marking. A recent investigation of bias in RCTs found that the lack of blinding in administration marking might cause a difference in the effect size (Ainsworth et al., 2015). In order to avoid marking bias, I covered the names of the schools and the anonymous code when I scored the creativity assessments. The reason why I chose to do this is because I did not want to even unconsciously disadvantage the comparison group with my marking. That would have resulted in finding the intervention I trialled as effective even if this was not the case. In other words, I tried to avoid type I error as it would have been called in statistics, which means I tried to avoid finding a false positive finding if this did not exist.

Inter-rater and intra-rater reliability were considered. Before even started the marking process, there was a moderation process. My two supervisors and I scored some assessments together. We discussed thoroughly responses and categorisation on the first activity and we compared our scores for the second creativity activity. That led to the creation of a scoring rubric which is presented later in this chapter. During the process, I had also regular meetings with them to discuss and compare our views of whether answers should be considered valid or how to interpret and categorise them.

Hence, there was an effort to ensure inter-rater reliability in order to succeed a more reliable marking. For the marking process, it was not feasible to constantly have two raters that would have been a good option particularly for the creativity

assessments. However, there were more than 1,500 questionnaires to be marked and there was no funding for a second rater to be hired. Second raters assisted at different parts of the process. The inter-rater reliability was examined and was found high (more than 0.9) after the adjustment of the scoring rubric of creativity which will be presented later in the chapter. When there was some disagreement between the second rater and myself, I had a discussion with the second rater and in some cases I adjusted my score, whilst in others I retained it.

8.1. Creativity: Activity 1

In both of the pre-test and post-test assessments, the first activity required the students to name as many uses of objects as possible. These replies, however, were not evaluated only based on the number of answers provided (fluency). This activity also aimed to evaluate flexibility and originality. Hence, there were three main indicators of creativity. Students could mention many answers about the uses and they can generate many ideas. However, creativity is not – and should not be – only related to the quantity of uses that somebody can generate, but also to the quality of these responses.

For these reasons, the two additional indicators were evaluated. The first indicator evaluated the creativity that the person demonstrated individually. In other words, even if the person produced many responses and suggested multiple uses for the objects, the uses suggested might have not been different from one another. Therefore, within the subject it was possible that more or less creativity was demonstrated based on the quality of the responses. The last indicator evaluated the originality or prevalence of the responses provided within the cohort overall. In that sense the creativity of the person was judged in comparison to their peers. In what follows, the three levels of data analysis which led to the marking of the first creativity activity are discussed. First, the grading method is described and then the technical aspect of grading is presented.

8.1.1. First level of analysis: Fluency

The first level of data analysis of the first activity included the fluency marking. For fluency, the students got a score represented by the number of the uses they wrote without almost any qualitative evaluation of the replies. I explicitly mentioned the word ‘almost’ in this definition of fluency because there are still some answers which can be considered invalid.

That level of analysis aimed at providing an indicator about the quantity of responses that the students produced. In that analysis level each answer was read and evaluated as valid or invalid. Almost all of the replies that the students provided were considered valid except a) the illegible answers b) answers which suggested the same use with the exact repetition of the same words or an exact synonym and c) answers which did not suggest an obvious use. It is crucial at this point to explain each of these categories in order to justify why these answers were considered invalid and provide some examples.

In the first case, the illegible answers were considered invalid. As the main investigator, I firstly identified illegible answers. However, then I asked advice from two peers – usually working at the university – of whom at least the one was a native speaker. If all the three of us could not read the answer and what was written, I considered the answer illegible. It has to be reported that there were a few cases that myself and another peer judged the answer illegible and the last peer managed to read the response.

In the second case, an answer was judged invalid if the student repeated the same use with the same exact words or an exact synonym (see Table 8.1.). The first type of repetition with the use of the same exact word is a straightforward judgment process. However, for the latter type of repetition with synonyms I tried to judge answers as invalid only when it was an exact synonym. The two words or two phrases written by the student were judged in order to be decided whether they were the same or different. If they were judged the same, then the second response was considered invalid. According to this guideline, if a student mentioned both ‘breaking’ and ‘snapping’ as uses of a pencil the one of the two answers were considered invalid, since these two words are considered to carry the same exact meaning. Even if the two uses had a slight difference, then both answers were considered valid. For this reason, for a few cases the advice of a native speaker was also asked in order to ensure whether two cases were exact synonyms since the dictionaries and an online translation were not always able to provide an accurate match between the two words or phrases written by the student.

Table 8.1. Examples of Responses and their judgement as valid or invalid.

	Reply A	Reply B	Judgment	Marks given
	X	Y (It is very likely that X is implied)	No repetition (both valid)	2
Example:	Writing	Books (the student is very likely to mean writing books or in books)		
	X	Z (exact synonym)	Repetition	1
Example:	Flat	Apartment	Only one is valid because both words have the same exact meaning.	

In the third case, an answer was judged invalid if it did not refer to a use even in an indirect way. This means that even if there was a slight possibility of the word or phrase to be considered a valid use, then this phrase was included in the fluency score. Nevertheless, if the answer was referring to a material or was self-referential to the object discussed and not its use, then the answer was considered invalid. One of the most common cases where the students provided invalid answers was because they seemed to interpret the question differently. For example, some students wrote replies which referred to the benefits of pencils instead of the uses of pencils, such as 'you can rub it out', 'they never run out like pens normally do' or 'they do not require a lid'. These replies emphasised the benefits of pencils and they had merit. However, they did not answer the question of the assessment and therefore they were not considered valid.

However, for this level of evaluation, the lack of some context and shared experience between children made the marking difficult. For example some children wrote 'Charlie' as a use of pencils. It was necessary to search for the response in order to find out that the students were referring to Charlie challenge which uses pencils. So the students mentioned something innovative, but I had to search for it in order to

identify that it is a valid and creative response. This is why the answers were judged carefully before concluding that they did not refer to any obvious use.

The responses which even in an indirect way suggested a use were considered valid because it was crucial not to disadvantage students who might have had some difficulty in writing. It was already a requirement for the students to have some basic literacy skills since they were required to write the responses. This means that the assessment included some construct irrelevance, since it did not measure only creativity but also basic literacy skills. However, there was an attempt to reduce construct irrelevance. For this reason, the students were not penalised if the use was not written in a clear way as long as there was even a hint that they mentioned a use. To be precise and demonstrate this point, Tables 8.2. and 8.3. present some of the answers written as uses of a pencil and bricks respectively and explain why answers were judged as valid or invalid.

Table 8.2. Examples of Student Responses for the Uses of Pencils and their scoring.

Student Responses for Uses of a Pencil	Judgment	Explanation	Marks given
Rubber	Valid	The pencil can be used to rub things with the rubber having on the top.	1
Rub it out	Invalid	The pencil can be rubbed out, but this is technically a use of a rubber not a use of a pencil.	0
Feeling	Valid	It might be the case that the pencil is used to be felt. For example, a mindfulness activity.	1
Paper	Invalid	It is a material. Pencils are usually used in order to write on the paper. However, there is	0

		not an implied use in this response.	
Break	Valid	The agent can break a pencil for different reasons for example as a stress reliever. Therefore, this is a use of a pencil.	1
Throw	Valid	This is a valid use because you might throw a pencil for different reasons. For example, somebody might throw a pencil as a game.	1
Drop	Valid	This is a valid use. There might be different reasons that somebody will drop a pencil. For example, somebody might drop a pencil to catch the attention of somebody.	1
Pencil case	Invalid	Not a use	0
Colouring pencils	Invalid	Self-referential to the object set by the activity	0
Wood	Invalid	This is a material. Pencils are made of wood. However, there is not a use implied in this response.	0
Lead/Graphite	Invalid	This is a material associated with pencils. However, there is not a use implied in none of these responses.	0

Books	Valid	The pencils can be used with books. When pencils and books are used together, then the use that is implied refers to the pencil. Thus, it might be a literacy issue here. The pencils can be used to write on a book or help a student to read the lines of a book.	1
School	Valid	The pencils can be used at school. This is a general use, but it might be an issue of literacy.	1
Learn	Valid	For the same reason that the word ‘school’ is valid, pencils can be used for learning.	1
Making	Valid	This is a general use. Pencils are used as a construction material. However, it might be a use of literacy.	1
Water	Invalid	All the types of material were judged as invalid responses.	0
Key/Locker	Valid	Both responses are valid because the pointy tip of the pencil could	0

		be used in order to open a locker.	
Glasses	Invalid	This response as a use for a pencil could imply many different things. For example, it could mean that pencils are used in order to support glasses that somebody is wearing. The student might have meant that the pencil builds the glasses or break glasses as a weapon or keep a window open. In order not to imply a use that was not actually mentioned. However, for consistency reason all the words which refer to materials are considered invalid.	0
Boxes Mail box Cardboard Marble jar	Invalid	There is no clear use implied. As in the case of a 'pencil case', it is only a container without a clear understanding of how a pencil can be used. It might be filling, or it might be opening a box, but there is	0

		not a suggestion of a use from this response.	
To open a box	Valid	There is a clear use.	1

Table 8.3. Examples of Invalid Responses for the Uses of Bricks.

Student Responses for Uses of a Brick	Judgment	Explanation	Marks given
Chicken and chips	Invalid	Even though food-related activities are acceptable as a response for a brick, simply mentioning food did not imply any use.	0
Sun	Invalid	This could not be a use for a brick.	0
Rainbow	Invalid	This does not imply even indirectly a use for a brick.	0
Key to Bravery	Invalid	There is no use stated here and it is not easy to understand what the student implied. Even though bravery would be rewarded in as abstractness, this was not evaluated in this activity.	0
Messy	Invalid	This is an adjective and there is no use suggested.	0

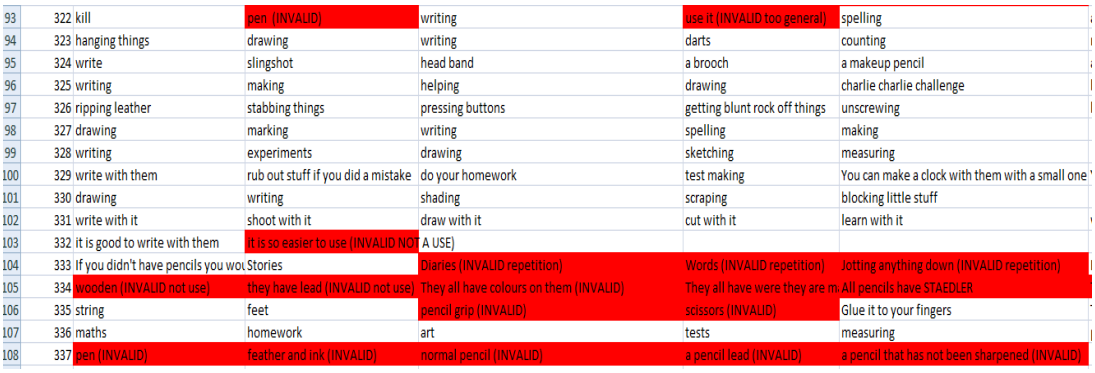
To summarise, for this first level of analysis the assistance of additional raters and native speakers were important. The evaluation of valid and invalid responses demonstrates the challenges of judging the content of the answers for identifying synonyms for a non-native speaker and the challenge of reading the handwriting of some of the students.

Concerning the marking, at the end of this level of analysis, all of the valid answers given by the students were counted. Each answer counted a single mark which represented the fluency score for that student. Each student was given a number as a score for fluency.

8.1.2. Fluency Analysis: Technicalities

All the responses of the students on the first activity of both pre-test and post-test were inserted as raw data in an excel spreadsheet. Each row of the spreadsheet represented a different student. Each cell represented a response. On the first level of analysis, any answer that the students provided was inserted in the spreadsheet. Screenshot 8.1. presents a part of the Excel spreadsheet at this first level of analysis. The cells with the invalid responses were marked red and all the rest valid responses were counted.

Screenshot 8.1. Fluency Analysis.



93	322	kill	pen (INVALID)	writing	use it (INVALID too general)	spelling
94	323	hanging things	drawing	writing	darts	counting
95	324	write	slingshot	head band	a brooch	a makeup pencil
96	325	writing	making	helping	drawing	charlie charlie challenge
97	326	ripping leather	stabbing things	pressing buttons	getting blunt rock off things	unscrewing
98	327	drawing	marking	writing	spelling	making
99	328	writing	experiments	drawing	sketching	measuring
100	329	write with them	rub out stuff if you did a mistake	do your homework	test making	You can make a clock with them with a small one
101	330	drawing	writing	shading	scraping	blocking little stuff
102	331	write with it	shoot with it	draw with it	cut with it	learn with it
103	332	it is good to write with them	it is so easier to use (INVALID NOT	A USE)		
104	333	If you didn't have pencils you wo	Stories	Diaries (INVALID repetition)	Words (INVALID repetition)	Jotting anything down (INVALID repetition)
105	334	wooden (INVALID not use)	they have lead (INVALID not use)	They all have colours on them (INVALID)	They all have were they are m	All pencils have STAEDLER
106	335	string	feet	pencil grip (INVALID)	scissors (INVALID)	Glue it to your fingers
107	336	maths	homework	art	tests	measuring
108	337	pen (INVALID)	feather and ink (INVALID)	normal pencil (INVALID)	a pencil lead (INVALID)	a pencil that has not been sharpened (INVALID)

8.1.3. Second level of analysis: Flexibility

Even though fluency referred to the number of responses produced by the students, the assessment also included qualitative indicators. Some responses had more merit than others and particularly it was noticed that within the same student creativity varied. The second level of analysis assessed the flexibility of the responses. The flexibility assessed how many different approaches each student suggested. For this evaluation, all the responses of the students were categorised based on their meaning (these categories can be found in the appendix 3a). This was a challenging process because

most of the times the student simply mentioned a word and the use of the object was implied.

To be more precise, for the quality of responses, it had less value if a student repeated the same type of responses, whilst it had more value if the student mentioned many uses which were distinctively different from each other. Flexibility examined the number of different uses each student mentioned.

It is important to clarify that the marking of flexibility categorised in the same category responses that they appeared similar. Examples of similar responses can be found in Table 8.4. In a sensitive measurement tool, some of these responses would have been attributed a different score based on different merit. Unfortunately, given the large number of responses and the fact that I was the only rater, the responses were not evaluated with such sensitivity.

It was common for the slight differentiated responses to be more humorous. Humour is an element which is evaluated by Torrance Test (Torrance, Ball & Shafter, 2008). Similarly, Davis (1999) included humour in the personality traits of creative people. Indeed, it has been found that the verbal creativity, which is the creativity mostly examined by this thesis, is associated with the presence of humour (Nusbaum, Silvia, Beaty, 2017). Furthermore, it has been found that there is some correlation between divergent thinking fluency and humour (Kellner & Benedek, 2017).

Table 8.4. Examples of Different Responses with similar content.

Responses	Slightly differentiated responses	Judgment
Weapon	Weapon (don't use it as a weapon)	The second response can be considered different than the first one. The second student can be considered humorous or sensible. What is included in the parenthesis seems like an ethical consideration from the student to avoid violence. Davis (1999) included the traits of emotional and ethical in list of the personality traits for creative people.

Build a house	Build a house for the homeless	The second response includes some sensitivity and some ideology, while the first one is only functional. Davis (1999) included this in the personality traits of the creative people. Creative people are empathetic and sensitive to the needs of others.
Eat it	Try to eat it if you are dumb	The second response includes some humour. The same use is stated but the student recognises that this is not a rational use.
Tower	Tower of respect	The second response includes an abstract concept. According to what is evaluated in the second activity, the abstractness is considered more creative. However, this activity does not measure abstractness and therefore this is not rewarded at this part of the assessment. Even though this might seem unfair, the assessment stayed focused and did not aim to assess too many elements at the same time. A grading focused on specific grading criteria was prioritised in order to keep the grading consistent.

It has to be recognised that it is a limitation of the measurement tool used for this research that did not acknowledge such a slight differentiation in the responses, which might reveal an additional merit. Similarly, some students wrote the uses as a story, but this type of writing was not rewarded. For example, a student wrote the uses of a brick as a story instead of phrases or words:

> With bricks I would dig a massive hole in my garden and make it 5 m deep. Then I would fill with dirt I see in bricks. Next I would get a ladder and throw it down the bricky room. I would make a hotel above the room and go inside with a lamp and a sleeping bag. Finally, I will survive the night underground with just a lamp and a water bottle.

Moreover, there were cases where the same word could be categorised in more than one uses. For this reason, there were a few criteria used in order to identify the use suggested by the students. The first criterion involved the examination of the words written before or after the response. For example, for the use of pencil some students mentioned 'science'. This word was categorised in the common use of the pencil, as writing, because 'science' was usually mentioned in a series of subjects in the responses, so the traditional use of pencils is implied. This is why 'science' was not categorised in the same category as experiments.

Also, if the word used by a student did not suggest a clear use, but other students explained the use, then the explanation provided by few students was used as a guide for the replies of the others. Thus, the explanation of the few students was used almost as a 'think aloud' protocol, giving access to what other students thought when they mentioned a specific word or phrase. For example, some students mentioned the word 'darts' as a use for pencil. This could suggest either a game or a weapon. However, because some students wrote 'play darts' and no student suggested that darts could have been a weapon, then all the responses related to the word darts were placed in the category sports. Similarly, the responses 'Shapes' and 'Rectangles' were categorised in the category writing for the bricks because there was another response 'Draw a shape' which suggested that the first responses might refer to drawing. Furthermore, for consistency reasons it was important for all the responses using the same word to be categorised in the same broad category. Thus, if there was no important reason to suggest otherwise, responses with the same word were categorised in the same category.

As it was explained at the beginning of the chapter, consistency of the marking was one of the most important aims of the marking process. Good assessments should be reliable. For this reason, some responses were categorised strictly in order to prioritise consistency. For example, even though 'punching bag' as a use of a brick

could go to the category ‘being aggressive towards the brick’, it was categorised in sports because the vocabulary suggested sports. Similarly, ‘Seat’ as a use of a brick goes in category 1 with responses such as ‘chair’ despite the fact that it could also be categorised in category 20. For consistency reasons, all the furniture was included in this category 1 of construction, whilst responses were included in category 20 only if no transformation of brick was suggested in the response.

In most of the cases for ambiguous responses the two previously mentioned criteria could shed light to the meaning of these responses. Therefore, the context of the previously mentioned words or the words that followed and the explanations that some students provided led to a better categorisation of the responses in these qualitative categories. In some of the cases the words were categorised in the category which seemed to be the closest related to what the student meant. For example, the word ‘laugh’ as a use for a pencil could probably go in many different categories. However, it was judged that it fits entertainment more than other categories. There were other words like this, such as the word ‘Internet’. Similarly, the word ‘jam’ for a pencil could even categorised as sound making or plugging, but it was judged that it probably fitted ‘plugging’ more than jamming. Similarly, the cages were categorised in category 1 instead of 2 and hospitals in category 1 instead of the category which refers to health look table 2 in the appendix 3a.

Nevertheless, these decisions could be to some extent arbitrary by the person who rates the assessments. Therefore, it can be argued that there is a level of subjectivity when these categories are concerned. This is indeed the case and it is something that I would like to be open about when the results are interpreted. Nevertheless, I did not expect that these could affect the results. There were more than 8,000 responses mentioned by the students which were categorised and there were only a few cases where the decision felt arbitrary to some extent.

It is important to be clarified that the cultural knowledge and general knowledge did not give more points in a creativity assessment. The construct evaluated in the assessment does not include any knowledge. In this sense, a student who wrote ‘building’ will get the same marks with a student who wrote ‘Big Ben’ as a use of a brick, because both students refer to the common use of a brick which is for construction. Even though the cultural knowledge was not rewarded by the marking, the assessor used cultural knowledge in order to interpret the responses of the students. For instance, the response ‘space’ as a use for a pencil is likely that it meant the

process of learning how to write in many primary schools where the teachers advice their pupils to use a pencil to keep a space between the words. Similarly, words such as 'mommy' as a use of a pencil were also categorised in the writing category because it is common for the students to draw their mum.

As it has already been explained, there was no consideration of merit of the responses. For example, if a brick is used as a 'punching bag' the person would probably break their arm or fingers. However, this response was still considered acceptable and put in the category, even though this is probably an unrealistic response.

Having categorised the responses, all the qualitative comments were turned into quantitative data. At this point, the number of different categories each student mentioned was measured. At the assessment of flexibility, mentioning the same category multiple times gave only one mark to the student's overall score. At this second level of analysis, the variation of responses provided by the students was graded. If a category was mentioned more than one time by the same student, it was deleted. Therefore, this revealed how many students mentioned each category. A table with number of students who mentioned each category can be found in the appendix 3b.

Finally, it should be clarified that the way flexibility was scored meant that students could score in flexibility an equal or lower mark than in fluency. Therefore, if all the answers of the students belonged to different categories, then a student got the same score for fluency and flexibility. For example students scored 5 in both fluency and flexibility only if all the five answers they provided belonged to different categories and suggested five distinctive uses. If some of the replies conceptually belonged to the same category, then the flexibility was scored lower than fluency. If every idea was distinct, then fluency and flexibility scores were the same.

8.1.4. Flexibility analysis: Technicalities

For the flexibility analysis, all the responses of the students were categorised. Each of these categories was given a unique code. All the categories with the unique codes and examples of responses which were categorised in each category can be found in the appendix.

Then, in the Excel of the raw data all the responses were recorded in the unique code of the category. All the invalid responses were categorised as zero. Therefore, the data after this recoding looked like Screenshot 8.2.

Screenshot 8.2. Data recoded into the codes of the unique categories.

10001	14	1	14	12
10002	1	18	5	5
10003	1	0	1	1
10004	1	11	4	10
10005	1	1	52	11
10006	1	11	1	9
10007	11	1	12	49
10008	12	81	5	1
10009	9	5	17	17
10010	1	5	54	12
10011	1	1	1	10
10012	1	1	12	76
10013	1	10	12	2
10014	1	5	19	
10015	6	20	38	1
10016	1	12	9	2
10017	1	5	9	47
10018	1	1	1	9
10019	11	12	6	17
10020	1	1	1	1

After this recoding, it became apparent that some of the students gave responses which belonged to different or the same categories. For example the student 10020 mentioned several responses which all were categorised in the category 1, the common use of the brick. On the other hand, the student 10008 gave responses which belonged to different categories.

Even though, this becomes easily observable in a screenshot, it was necessary to identify a way to measure the flexibility score for each student, which meant to measure how many different category codes there are in the responses of each student. For this reason, I created a new excel. In this Excel (see a part of this Excel as an example in Screenshot 8.3), each student was a column, whilst each raw represented a response. The maximum number of rows I used was based on the fluency that the students had, so there were 39 rows for the pre-test. Below these rows, there was each raw for each category which basically measured whether the student in the same column had mentioned these categories.

In other words, I used the option ‘CountIf’ to identify whether a student mentioned a category, and this was combined with the option if. The option if turned all the number bigger than 1 to 1. For example, for the student in the column C when the category 8 is concerned I wrote in the Excel

IF(COUNTIF(C2:C40,8)<>0;1;0)

Therefore, in the calculation below the data (look screenshot 8.4.) the number 1 meant that this category was mentioned at least once whilst 0 meant that the student did not gave a response related to a category. Then, I added the number of different categories which were coded as 1, for each student in order to get their flexibility score.

Screenshot 8.3. Number of Responses and Categories mentioned by students (Unique students codes on the first raw with the unique categories codes that each student mentioned).

B	C	D	E	F	G	H	I
	101	102	103	104	105	106	107
item 1	1	1	23	1	6	1	
item 2	1	44	3	1	1	1	
item 3	1	25	1	0	39	5	
item 4	1	2	1	1	14	14	
item 5	1	1	1	85	1	2	
item 6	0	1	1		5	17	
item 7		1	2			25	
item 8			1			10	
item 9							
item 10							

Screenshot 8.4. Turning the fluency scores into flexibility scores. (In the column on the left you can find the unique code of the category. This table identifies whether a category existed (1) or not (0) for the student which is in the same column).

1	1	1	1	1	1	1	0
2	0	1	1	0	0	1	0
3	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	1	1	0
6	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	1	0
11	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0
14	0	0	0	0	1	1	0
15	0	0	0	0	0	0	0

8.1.5. Third level of analysis: Prevalence

The third level of analysis included the evaluation of the prevalence of responses. For this purpose, the answers of each student were compared to the answers of the other students within the cohort. Assessments which report the performance of students in comparison to the performance of other students within a group are called norm-referenced (Koretz, 2006, p. 50). In assessments like this, an original answer in one cohort may not be an original answer in another cohort. Therefore, in the assessment

used in this research, scoring prevalence as a third indicator of creativity in activity 1 can be considered an item which attributed scores in a norm-referenced manner.

At this third level of analysis, there was a calculation of a score for each of the categories created in the previous level of analysis which evaluated flexibility. That score represented the number of people who mentioned this category within the overall research sample who took the assessment. Therefore, each category which already had a unique code for the flexibility analysis was not attributed a Frequency Code. Even though each category had a unique Code for the prevalence analysis, the Frequency Code could be the same for more than one category. For example, a Frequency Code 19 would mean that a category was mentioned by 19 people in the overall cohort who sat the assessment. If another category was also mentioned by 19 people, then that category would still receive a Frequency Code of 19.

As a result, each response of the students could receive a Frequency Code based on the prevalence of this response amongst all of the students in the research sample. It has to be clarified that as in the flexibility score each category was counted only once. These had two consequences. If a student mentioned twice the same unique category, that was linked only once to a Frequency Code. Hence, there was no Frequency Code higher than the number of students in the sample.

At the last stage of this analysis, the aim was to attribute a Prevalence score to the students. There was a challenge to decide on how this score would be calculated. If the Frequency Codes were used and aggregated, then the lower the score that a student got the more creative that student would be. However, what usually happens in assessments and in the previous two levels of analysis is that the more answers students would suggest, then the students would be considered more creative.

That would be a big contradiction in the measurement of creativity. However, it was desirable that the higher scores were given to the most creative students. Hence, there was a reversion of the scores and recode of the categories of the students mentioned. In order for this to be achieved, it has to be reminded that there was no Frequency code higher than the N (number of students in the sample). This is due to the fact that each category was counted only once even if mentioned more than once by the student. This means that P_k results from the flexibility and not the fluency score. Hence, in N is overall number of the students who sat the assessment, then the Prevalence Code for each category was calculated as

$$\text{Prevalence Code} = N - \text{Frequency Code}$$

The calculation of prevalence code for each unique category based on its Frequency Code can be found in the appendix. What should be clarified at this point are the two different ways to measure prevalence that they were established. The first measurement of prevalence was based on the overall score and it was called Prevalence Sum. The prevalence sum is the score given to the students and it is calculated as the sum of all the prevalence codes of the responses of the student. As a result student who score high in this variable should score both have mentioned many answers and answers of categories with high prevalence scores.

However, this type of calculation could offer an advantage to the students who mentioned more categories, because they were be more variables to be added. For this reason, there was also a different calculation of prevalence score for the student. That second variable was called Maximum Value and was equivalent with the category with the highest prevalence score in the responses of the students. Therefore, the score of students for this variable would be dependent on the rarest category they mentioned in their responses.

The maximum value was also calculated because a student might not have received a high score overall, but they could have offered an innovative answer. For this reason, I decided to correlate and examine both of these indicators in order to choose which one is the most appropriate to be used. Only one of the two variables would count as a prevalence indicator in the overall score of creativity and the decision is presented in the results section.

Before, discussing the prevalence score, an adjustment was necessary. Due to the fact that N was bigger in pre-test than post-test, that slightly disadvantaged the group in the post test. To be more precise, if only one student gave a unique response in the pre-test that student scored 816 in the pre-test. A student who gave a unique response in the post-test scored 737. There is an obvious disadvantage in the second group. In order to make the two scores equivalent despite the fact the difference in the sample size in pre-test and post-test, the final prevalence scores of the students (both sum and maximum value) were calculated using the formula below

$$Final\ Prevalence\ Score_{pre-test} = \frac{Prevalence\ Score}{817}$$

$$Final\ Prevalence\ Score_{post-test} = \frac{Prevalence\ Score}{738}$$

Similarly, the maximum value was adjusted

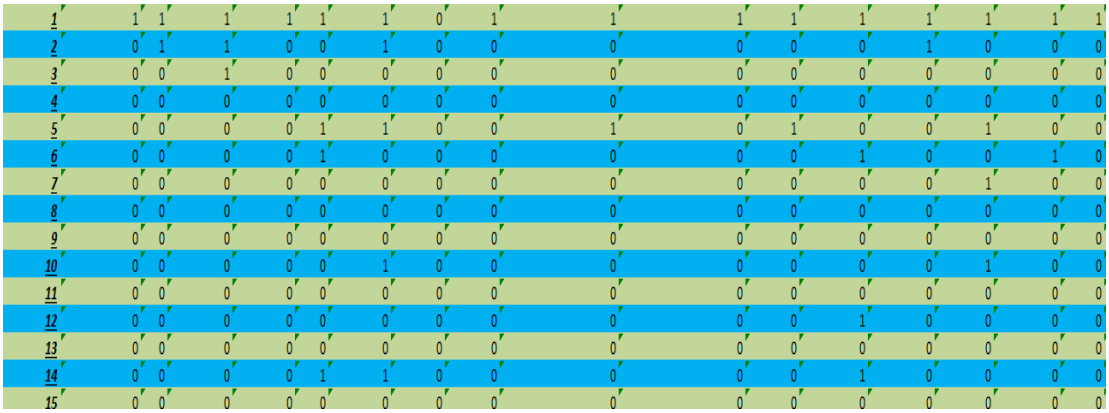
$$Maximum\ Value_{pre-test} = \frac{Maximum\ Value\ Prevalence\ Score}{817}$$

$$Maximum\ Value_{post-test} = \frac{Maximum\ Value\ Prevalence\ Score}{738}$$

8.1.6. Prevalence analysis: Technicalities

The Excel that I used for flexibility scores was the excellent basis for the calculation of the Prevalence. Screenshot 8.5 presents a part of the Excel with the form of data which were used in order to calculate the unique frequency code for each category. First of all, each category was already located in each row and it was counted only once for each student. Therefore, I calculated the sum for each raw and that gave me the Frequency Code for each category.

Screenshot 8.5. Example of the Data used for the Calculation of Unique Frequency Code for Each Category.



Then, I calculated the prevalence code for each category as it can be found in the appendix 3c. I copied and pasted all the data in a new Excel spreadsheet. On the right column there was the unique code for all the categories and on the first raw the unique code for each student. Then, I recoded all the '1' which stood for category 1 with the Prevalence Code of that category. Similarly, I recoded all the 1 which stood for the

category 2 with the Prevalence Code for that category and I did the same for all the categories as presented in Screenshot 8.6.

Screenshot 8.6. Recoding of Flexibility Score of each category into Prevalence Code.

	9001	9002	9003	9004	9005	9006	9007	9008	9009	9010	9011
1	48	48	48	48	0	48	48	48	48	48	48
2	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	635	0	0	0	0	0	0	0
8	0	0	0	0	0	0	697	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	687
14	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	645	0	0	0	0	0	0	0
18	0	0	0	0	0	0	673	0	0	0	0

This recoding created a new database on which it was easy to calculate the Prevalence Sum score and the Maximum Value for each student (see screenshot 8.7). Finally, these scores were adjusted by dividing with the overall N, because as mentioned earlier the overall sample in pre-test was different from the one in the post-test.

Screenshot 8.7. Calculation of Prevalence Sum and Maximum Value from the Prevalence Codes.

46	70	0	0	0	0	0	0	0	0	0	0
47	72	0	0	0	0	0	0	0	0	0	0
48	74	0	0	0	0	0	0	0	0	0	0
49	75	0	0	0	0	0	0	0	0	0	0
50	78	0	0	0	0	0	0	0	0	0	0
51	81	0	0	0	0	0	0	0	0	0	0
52	85	0	0	0	0	0	0	0	0	0	0
53	86	0	0	0	0	0	0	0	0	0	0
54	93	0	0	0	0	0	0	0	0	0	0
55	100	0	0	0	0	0	0	0	0	0	0
56	SUM	48	48	48	1328	0	48	1418	48	48	1498
57	MAX	48	48	48	645	0	48	697	48	48	728

8.2. Creativity: Activity 2

In the activity 2 is based on an activity of the Torrance test and how it was used in a previously published thesis (Shaheen, 2010). The scoring rubric of Torrance test was adjusted in both activities used by this research. The rubric for both activities can be found in the appendix 3d. The reason why this happened will be explained in this

section. For the same reason that some other variables, such as humour, were not included in order not to make the judgments subjective, the scoring rubric of Torrance was adjusted in order to be marked in a consistent way and to improve interrater reliability.

From this activity, the abstractness of the title and the premature closure of the students were assessed. The activity and the marking scheme (look appendix for the exact marking scheme used) are based on Torrance Test of Creativity Thinking (Torrance, Ball & Safter, 2008). However, I adjusted the marking scheme to be more precise. Furthermore, since the assessment of this activity is inter-subjective, and it can depend on the judgment of the assessor, I decided to use a smaller scale (0-2). I espouse that the more numbers this scale includes, the more fluctuation we could have in the scores. As this is a subjective evaluation, I decided to restrict the scale to avoid the variance and the potential arbitrariness.

The assessment of the resistance to premature closure can be challenging because it involves the evaluation of the shape of the picture drawn. To be more precise, the picture was scored with 0 if the figure is closed in one of the quickest ways, or the student wrote a letter(s) of the alphabet or number(s). This score was also given if the student closed the shape with one of the quickest ways and added details within the closed figure. The picture was scored with 1 when details were added outside of the enclosure. Finally, the picture was scored with 2 if there is no closure (the shape is open) or the shape was closed with the use of irregular lines as part of the picture. The way the marking scheme was phrased was different to the Torrance test. The aim was to make the marking scheme phrased in the simplest possible ways. Furthermore, the marking scheme in Torrance test (Torrance, Ball & Safter, 2008) created a big ambiguity when there was a reference to the drawing which score 2:

> Closure is never completed is completed with irregular lines which form part of the picture rather than with straight lines or simple curved lines. (p.13)

The ambiguity concerned the first case. The test should clarify that the drawing is not completed but it is included in the drawing somehow. Otherwise, all the students who did not draw anything and left the item blank would have scored 2, since the drawing would be incomplete. Thus, I decided to change the phrasing in the marking scheme concerning this item.

This activity scored the resistance to premature closure and the abstractness of the title, which are two characteristics that creative people have. Initially, the peer raters and I tried to use exactly the Torrance Test scoring rubric. However, there were difficulties and weaknesses identified in this rubric. First, for the resistance to premature closure the test gives 0 for closed shapes, but 2 marks for a shape that is closed but considered to be closed with irregular lines. Therefore, what was considered irregular lines were sometimes questionable and this could cause a big variation between raters. If the lines were not considered irregular a student could score 0, whilst the same student could score 2 if marked by a different rater who considered the marks irregular. Torrance test might address this inconsistency with training of the raters and might mediate this problem because of the big number of pictures evaluated in this task. When this rubric was used for the piloting of this thesis, the inter-rater consistency was low. Therefore, a new rubric was created. Therefore, the marking related to combination of pictures was not applicable in this trial, because I included only one image given the concentration span of the age of the participating students.

Concerning the abstractness of the title, the scoring guidelines of Torrance test were also adjusted. Specifically, according to the guidelines the students score 0 if they state the obvious, the students score 1 mark if the title is simple descriptive with a modifier, such as ‘dancing cat’. However, the students score 2 marks if the title is imaginative and the modifier goes beyond concrete, such as ‘the dog named king’. Finally, an abstract title gets 3 marks. However, when there is a judgment between 1 and 2. First of all, the student might not have drawn something effectively and therefore something might not appear obvious from the picture. In that sense students who draw better might have been disadvantaged.

For the content validity to be achieved the activity should measure exactly what it states that it measures. The initial rubric gave additional marks for students who mention objects which do not exist. However, I adjusted the rubric in order to measure only the abstractness of the title independently of the picture. In this way the items of the assessment are kept independently. Furthermore, even though the importance of the students mentioning a title of an object that does not exist was recognised, this would still include many problems since it does not necessarily imply that it is more abstract. For fairness reasons, in the marking process it would be problematic to distinguish when students imagined an object or have seen it somewhere. For example, students might have seen an imaginary object to a TV programme or a book. It would have been

impossible for the researchers to identify all the possible influences and distinguish between demonstrating original imagination and reproducing of somebody else's imaginative ideas. Thus, these were included in the same category.

Consequently, the abstract titles got the highest marks. Furthermore, if the title implied the existence of a story, the title was also considered as involving a high level of synthesis and organisation. This characteristic is associated with creativity and therefore the title got high grades. It might appear bizarre that the concrete title gets more points than the simplistic title, which is graded with 0, when the construct rewarded is abstractness of title. However, as Torrance, Ball and Safter (2008) reported, the abstractness of title includes initially the individuals' ability to synthesise their thinking and in the highest level to capture the essence of information involved.

Even though for the 10% of the first data marking two more raters were used, it has to be clarified that one of the most important things was for me to be consistent. Since all the sheets were marked by me, it was important for me to consistent with myself (intra-rater reliability). Even if I was strict or lenient, since there were no other raters, the goal was my own consistency. To avoid conscious or unconscious bias based on whether the student is in the intervention or the comparison group, I marked the creativity activities in a blind way. When I was marking, I did not know whether a student was in the comparison or intervention group.

8.3. Scoring Process

In this chapter so far, the scoring process for each of the two activities of the assessment is described. However, the research question of this thesis examines the impact of the P4C programme on creativity overall. In the previous chapter, the way that the critical thinking overall score was calculated was described.

Despite the fact that the calculation of the critical thinking score overall was pre-decided, this was not the case for the calculation of creativity score. As it will later be discussed, there was existing literature which suggested that some of the evaluated aspects were highly correlated. As a result, simply summing of all the aspects would create unequal weighting for some domains of creativity which would be highly correlated with each other.

Therefore, I decided to create the formula of evaluating the creativity overall score after having the data. First, I would examine the correlation between the various aspects of creativity and based on this finding I would decide whether I should include

all the aspects or exclude some aspects which were highly correlated with others. Consequently, the way that creativity overall was calculated will be presented in the results chapter, since it was based on the specific data I collected.

9. Pilot Study

For the investigation of the second and third research question of this study, a comparative evaluation study was conducted. Before distributing the assessments for this large-scale project, it was necessary to conduct a pilot study. The two parallel forms used in the piloting can be found in the Appendix 4.a. and they have some differences from the finally used assessments. This chapter highlights the basic aims of conducting a pilot study and explains how each of these objectives was satisfied. The sampling method and the sample characteristics for the pilot study are cited. Moreover, the process followed for the piloting is described and the reason why I decided not to be present in the piloting is explained. Finally, the feedback is presented and the modifications of the items after the piloting are discussed.

9.1. The aims of conducting a pilot study for the measurement tools

Before conducting the actual experiment, it was necessary to pilot the measurement tools. In this section, the reasons for conducting a piloting are analysed. For this specific research, it was extremely important to pilot the measurement tools since the measurement tools were used for the first time and constructed for the purposes and the sample characteristics of this study. Therefore, there were no prior indicators concerning their reliability and validity. In the previous chapter, all the processes followed for the construction of the measurement tools was discussed. However, before distributing the test to all the schools of the study it was necessary to identify whether these tools were indeed appropriate for this age group, whether the students understood the instructions and to explore the items difficulty and discrimination.

9.2. The sample

In order to ensure that the questions were suitable for the students' in the study, the assessments were piloted in a school with similar characteristics to the targeted sample. A school in southern England volunteered to help and both test forms were administered in two separate classrooms. The specific school had been implementing P4C sessions for the last four years. In other words, the specific Year 5 students were students which participated in P4C sessions since they started the primary school. All the schools in the actual study were also located in England, so the location of the school in the pilot study matched the location of the participants of the actual study. It is important that both forms in the piloting were administered towards the end of the

school year, so the students were about to finish Year 5. If maturation plays a role in the performance of the students, then the students in the pilot study have an advantage compared to the study participants who completed the pre-test form at the beginning of Year 5.

Additionally, although I had not reached any conclusion as to whether P4C helps the development of critical thinking and creativity, I could not exclude the possibility. Therefore, I chose to pilot the measurement instruments with students who were likely to have developed critical thinking and creativity because of longitudinal P4C implementation. This decision was due to my aspiration to exclude the plausibility of thinking problems to be answered correctly by all the students in the actual study. By implementing the measurement tools at the end of the school year and with students participating in P4C sessions for many years, I accepted that they potentially have increased critical thinking and creativity. To summarise, my piloting was based on the hypothesis that the thinking skills (creativity and critical thinking) in the piloting group were higher or equal to the average of the students in the actual study.

$$CT_{\text{piloting group}} + Cr_{\text{piloting group}} \geq CT_{\text{actual sample}} + Cr_{\text{actual sample}}$$

Based on this hypothesis, there were two assumptions when the piloting school was selected:

a) If an item was taken wrongly by all the students, then it would be considered too difficult. If the students could not reply correctly to a question, this would entail that this item is not appropriate for the specific age and targeted group. As a result, that item should be either modified or excluded by the measurement tool.

b) If an item was answered correctly by more than half of the students, it was appropriate to be included in the final test. However, I decided to exclude the items answered correctly by all the students because items responded wrongly by some students are desirable, since that distinguishes the students between low and high performing groups. The group of students, who volunteered for the piloting, has similar or more developed thinking skills than a typical group. As a result, if

items are taken wrongly by some of the students, then in another group in the actual study there will be also a proportion of students who will take this item wrong and therefore the item will discriminate the students and it should be included. However, if P4C increases thinking skills and the thinking skills of these students are indeed developed, an item judged easy by this group of students, it would not be evenly easy for a different group of students.

Nevertheless, my hypothesis could have been wrong with P4C having a negative impact on critical thinking and creativity and therefore the measured thinking skills in the study will be higher. The reverse hypothesis could be summarised

$$CT_{piloting\ group} + Cr_{piloting\ group} \leq CT_{actual\ sample} + Cr_{actual\ sample}$$

If this hypothesis was accepted, then it was likely that the control group in the study will perform higher than the piloting group. Therefore, the scores of the control group could have been characterised by 'ceiling effect' with students responding to all the items right.

9.3. Administration process

The pilot was conducted in the same way as the one to be followed in the actual study. The school replied to an e-mail confirming participation. The P4C coordinator in the school informed about the precise number of Year 5 students. For each of the students a pre-test and a post-test were sent by post. In the envelope, the assessment forms, teachers' sheets (one for the pre-test and one for the post-test) and a sheet with the detailed administration instructions were enclosed. I provided the school with detailed administration guide to achieve a standardised process. In the envelope, there was also a pre-paid envelope, so the forms could be returned to the School of Education at Durham University. After I analysed the data I provided a cohort report with results deriving from both forms. Even though the aim of the pilot did not involve tracking the individual or the cohort performance, feedback was provided to the school which kindly co-operated.

9.4. Grading of creativity activities

The most important function of pilot regarding the creativity items were the pilot of the grading and coding system. I have already explained in the previous chapter how these two activities were graded. The methodology of the grading for the creativity activities was mostly developed during the pilot study and finalised during the grading of the first questionnaires of the actual trial.

The pilot study helped me decide on the way to grade these activities and take decisions to develop my consistency as a grader. It also revealed the inconsistencies in the marking process and the need for the development of a new grading system which was described in the previous chapter.

9.5. Item analysis of Critical Thinking Assessments

Concerning the thinking problems, which are multiple-choice, I searched for specific indicators which I judged that they could help me understand to what extent the tools were successful.

9.5.1. Item difficulty and item discrimination

There are two main theories for tests the last two centuries. The first is the Classical Test Theory and the other is the Item Response Theory. Classical Test Theory claims that the observed test scores are a combination of the true score and measurement error (DeVallis, 2006; Koretz, 2006). The true score is the average score that the person would obtain if the performance was measured repeatedly by similar assessments - assuming that there is no practice effect with the person becoming better because of getting used to the assessments (Cronbach, 1961, p.129). However, in the case of this pilot study it was not possible to calculate the true score because of the lack of repeated measurement tools.

Instead a Rasch model approach was used. The Rasch model primarily espouses that the score which can be attributed to a student depends on the student ability and on the difficulty of the items (Magno, 2009). In this analysis, there is a consideration of the item difficulty and item discrimination. According to the Rasch Model students' ability, item difficulty and discrimination are measured in the same scale. Item difficulty as the name suggests is the level of difficulty that one of the constructed thinking problems might have and it is calculated by the proportion of students who got the item wrong. The item discrimination refers to the extent that 'an item differentiates correctly among test takers in the behaviour that the test is designed

to measure' (Anastasi, 1988, p. 210). In other words, it refers to the extent that an item distinguishes effectively the high performing students of the low performing concerning their performance on the measured traits.

Specifically, the correct answers were scored with 1 and the wrong with 0, thus the items were scored dichotomously either right or wrong. After the administration and scoring of the test, it is possible to estimate the item difficulty and the item discrimination. Item difficulty in dichotomous item score can be calculated by calculating the mean of each item in SPSS (Frequencies>Descriptive statistics). Thus, items which have mean = 1 (when the label 1 means that the student has taken the item correct), are the items which have been answered correctly by all the students. From these means the item facility can be estimated, so an item which has facility 1 has been answered correctly from everyone, so it has 0 difficulty. The mean of each item represents its facility. It is really easy with a simple subtraction to calculate the item difficulty (1- mean). The item difficulty and discrimination for both of the problems are presented (Table 9.1.-9.4.).

Table 9.1. Item Difficulty for Form 1

Thinking Problem 1: Does James ride a bicycle?	0.36
Thinking Problem 2: Who do you believe?	0.44
Thinking Problem 3: The meeting	0.17
Thinking Problem 4: Listening to classical music	0.31
Thinking Problem 5: Two friends were talking	0.25
Thinking Problem 6: The weather	0.92

Table 9.2. Item Difficulty for Form 2

Thinking Problem 1: Does your brother learn the guitar?	0.36
Thinking Problem 2: Who do you believe?	0.26
Thinking Problem 3: The road	0.20
Thinking Problem 4: Pocket money	0.29
Thinking Problem 5: An announcement	0.26
Thinking Problem 6: For the end...let's eat a cake!	0.31

The item difficulty is an extremely important factor. If an item is answered correctly by all the students, it takes some space in the assessment form, it requires time and effort from the students to be completed, but it does not provide any additional about pupils' ability. Therefore, the above analysis was important, because it confirmed that none of the items was too easy, since there was no item answered correctly by everyone. These values were interpreted according to the purpose of the assessment. The purpose of this assessment was not to rank the participants or to select the higher performing participants. On the contrary, the purpose of the assessment was to evaluate to what extent students were creative and developed the skill of critical thinking. In other words, this assessment could be perceived as an assessment of mastery of the critical thinking skill for group comparisons. According to Anastasi (1988, p.210) the item difficulty can be interpreted according to the use of the tests and particularly recommended mastery skill tests to have items with difficulty around 0.80.

Based on this recommendation, it could be argued that the items were too easy. However, I decided not to change them, because if my hypothesis was correct, with P4C leading to improvement of thinking skills and creativity, then these students were a cohort with more developed thinking skills than an average group, as they have been involved in P4C the last 4 years. Furthermore, since both forms were administered towards the end of the school year students were more mature than the participants in the trial. The problem 'Let's eat a cake!' should have been parallel for 'The Weather' problem since this was the respective problem of evaluating the problem-solving skill in the other form. Nevertheless, the problem 'Let's eat a cake' was too easy when compared to the problem 'The Weather'. Therefore, the first was removed. Two items of reasoning were included in both forms (thinking problems 3 and 4) in order to have reasoning of different difficulty. As I aimed when I designed the assessments, the first reasoning problem was more difficult than the second one (Table 9.1 and Table 9.2).

The discrimination of the test items is discussed in Item Response Theory in 2 or 3 parameters model (Sick, 2008). An item has a good discrimination when all or some of the high scoring students get it right, but low scoring students or almost all of the low scoring students answer wrong. On the contrary, it has poor discrimination when equally high and low scoring students get the item right and it has negative discrimination when solely low scoring students and not high scoring get the item right. About the discrimination for dichotomously scored items for normal distribution Pearson correlation in SPSS is done. Even though Pearson correlation has been used to

reveal the item discrimination, I do not report the statistical significance. In other words, I do not discuss whether this correlation has been found statistical significance. The statistical significance testing is based on the assumptions of randomisation in the sampling method (Gorard & Gorard, 2016) and hence it was not appropriate for this case. However, this process allowed me to identify items with low discrimination.

Table 9.3. Item Discrimination for Form 1

Thinking Problem 1: Does James ride a bicycle?	0.392
Thinking Problem 2: Who do you believe?	0.729
Thinking Problem 3: The meeting	0.443
Thinking Problem 4: Listening to classical music	0.516
Thinking Problem 5: Two friends were talking	0.421
Thinking Problem 6: The weather	-0.251

Table 9.4. Item Discrimination for Form 2

Thinking Problem 1: Does your brother learn the guitar?	0.539
Thinking Problem 2: Who do you believe?	0.485
Thinking Problem 3: The road	0.388
Thinking Problem 4: Pocket money	0.484
Thinking Problem 5: An announcement	0.485
Thinking Problem 6: For the end...let's eat a cake!	0.578

Concerning the specific item correlation with the overall performance of the students in the critical thinking test the desirable correlations were found. A correlation which is negative entails that students of low performance get the question right. This might be due of guessing or potentially construct irrelevance and it is apparent that none of them are desirable for a reliable and valid test. However, the number of students was low for any conclusion. These correlations only provided some indicators about the item discrimination. Furthermore, a test of a multi-facet construct like this, low correlations were expected because each item presented different information about the performance of the student on a different task and aspects of the construct.

Anastasi (1988, p. 211) recommended that item discrimination is not a useful indicator for a criterion-referenced mastery skill. In other words, since this assessment examined whether and to what extent students are critical and creative, the item difficulty was not examined. Instead of the item discrimination, Anastasi suggested the examination of criterion validity between the piloted assessment and a criterion assessment. However, in the area of critical thinking - as I hope it has already become apparent to the reader - there is not a gold standard assessment to measure the specific two thinking skills for students of this age. Thus, the criterion validation by correlating this assessment with an external criterion was not a feasible option.

9.5.2. Missing Data

Missing data was also examined. If the students left some responses blank, this would indicate that an item was difficult or less interesting. Furthermore, if this item happened to be at the end of the assessment, it could mean that there was insufficient time for the test to be completed. All the students replied to all the questions. The pilot study did not have any missing data and therefore it did not provide indicators for issues such as the aforementioned.

9.5.3. Pattern of correct and wrong answers

One of the things that were revealed though was the correct answers pattern. The correct answers of the test with a multiple choice cannot be always the "a" or the "c" answers. Thus, it has been an effort to balance the pattern of the correct answers. For the critical thinking test which had multiple choices item, I realised better the correct response pattern when I was correcting the forms.

The patterns for form A was: 1-C, 2-A, 3-B, 4-C, 5-C, 6-A

The pattern for form B was:1-C, 2-A, 3-B, 4-C, 5-C, 6-C

Option C appeared as the correct answer for most of the times. However, I did not want to keep a fully balanced pattern with each letter to be correct for 2 times (2 times · 3 letters for 6 problems) because key balancing might lead to predictability of the correct answers, testwiseness and guessing (Bar-Hillel & Attali, 2002). However, I considered the pattern of correct options for the assessments used in the trial.

Concerning the quality of wrong options which were used to distract (distractors) some of the students from providing the correct answer, there was an

analysis of the answers provided for all the thinking problems separately. Specifically, charts were created for each of the problems in order to shed light in the possibility of having a misleading distractor, which confuses more students than usually (see appendix 4.b.). Even though the sample was small, the answers of the students in each thinking problem show that the two distractors are equally misleading, but generally students were able to identify the correct answer.

9.6. Feedback

I considered whether I should leave blank space for the students to write comments at the end of the test. However, when there are only 36 students, it is possible that the productive comments will be just a few. For this reason, I decided not to take much time from the school which offered the help by asking also to provide qualitative feedback by the students in a written format but asked the teacher to search for oral feedback. Some students decided to write their comments on the questionnaire and the P4C coordinator of the school successfully kept a record of the students' comments. These comments provided a good insight on how the target group perceives the forms. Most of the comments were provided by the person who administered the test. A few comments were also written on the survey forms. The comments helped to identify omissions or missing information (F8) which led to some rephrasement of the thinking problems.

9.6.1. Thinking Problem: Does James ride a bicycle?

'If he rode a bicycle he would take care of his bikes, therefore I don't believe he rides a bicycle'.

This student believed that option B is the correct answer. The question to be set is whether the B option is too misleading. After this comment I examined the distractors. The option B was equally believable with the option A, but the majority of the students were able to identify the correct answer. Consequently, even though the distractor B is plausible, it was not judged too misleading. According to the results of pilot study, the students were able to judge that the information is not sufficient to lead to a solid conclusion and therefore they chose the option C.

9.6.2. Thinking Problem: Who do you believe? (Form 1)

'It doesn't tell us if she's going to be driving on the weekend'.

'It doesn't say if she's walking or driving'

'It doesn't say what happens on the weekend'.

'It depends which way she wants to go'.

The two comments about the weekend provided an insight of the item that I had not thought of. Both of the people pass the roads when they return from work - probably on weekdays. If I had an option 'none', then the justification about the weekend would be excellent. However, I thought that I might have had to specify that Nadia arrives in the city 'on Tuesday' and she wants to 'drive' (and not walk- as the third students says).Concerning the last comment about the way she wants to go, it is unnecessary to add information since it is already written in the problem that she is interested in the traffic on the Shaftesbury Avenue.

9.6.3. Thinking Problem: Listening to classical music

'Maybe she is at home but not in her room, maybe she's in the garden that is why she can't hear him'.

This is a comment which reveals the process of thinking for the student who finds the correct answer. However, it is not constructive or leads to any change in the test.

9.6.4. Thinking Problem: Two friends were talking

'*How does he know that every person drank orange juice? He couldn't possibly know that'.*

'Maybe she forgot she drank orange juice'.

I was impressed by the first comment. It is a critical comment and this type of thinking is what the test I constructed tries to investigate. At this point, it becomes obvious that multiple-choice items cannot capture all the alternative types of thinking and might be restricting. Nevertheless, the students were asked to take for granted whatever information is given and make a judgment based on the given information. Therefore, unfortunately they had to accept that Steve knew that everybody drank a juice. For example, there might have been a toast with all the people and it became obvious that

everybody was holding a glass with orange juice or Steve might have been serving the drinks all night.

There was no change made based on the second comment. I thought that mentioning that Charlotte has a fabulous memory would have been more confusing for the majority of the students who will not think the possibility of Charlotte forgetting the fact that she drank orange juice.

9.6.5. Thinking problem: Who do you believe? (Form 2)

'What is green tea'?

'Tea has water in it but that green tea stuff may not be good for headaches'.

The further explanations for the tea were not judged crucial for the thinking problem.

9.6.6. Thinking problem: The road

'*Maybe it rained but someone cleaned the road'.*

I examined to what extent the students were confused by the C distractor and it was not found particularly misleading.

9.6.7. Other comments

There were also other student comments that were transferred by the teacher. The comment was about the appearance of the questionnaire (F9). 'A student asked why you were using "" for dialogue. He said that in school they are used to the quotation dash – to indicate dialogue [...]. Some also wondered why there was no cloud with instructions on the last page'. For this reason, a thinking cloud was added on the last page of the questionnaire.

9.6.8. Teacher Comments

I also welcomed the feedback by the teacher. I was not present in the administration process and therefore I requested analytic feedback by the teacher who assisted with the pilot study. Even though it might be argued that in the first administration, I should be presented, I firmly believe that the pilot study should follow the exact same process as the actual study. 'As Oppenheim remarks, everything about the questionnaire should be piloted; nothing should be excluded, not even the type face or the quality of the paper' (Cohen, Manion & Morrison, 2007). For this reason, despite the fact that the

pilot study showed that adjustments should have been made in the assessment forms, the same administration process was followed exactly as the trial. I decided to post the assessment forms in order to follow the exact same process as the actual study.

In the teacher sheet for each form there was a question which asked the teacher whether there were words that the students did not know. In spite of having previously examined the readability scores, this questioned aimed to further explore to what extent the carrier language was appropriate for this age group. Carrier language is the language which is used to set the task (Chartered Institute of Educational Assessors, 2008). In other words, carrier language is the question which is set, and it looks for the answer. If the question is not explicitly set, then the answer is more difficult to be given. In high readability items the variable which is measured is not only the mathematical ability, but also other variables, such as the reading ability (Hewitt & Homan, 2003). The phrase that the students did not know according to the teacher comment was only one (F5); 'take for granted'. This means that the Readability test used to reassure the appropriateness of the language according to the age of the student was successful.

After the administration process, I contacted the teacher to ask further feedback. She kindly responded to a short questionnaire I sent to her. The questions in the questionnaire aimed to cover important functions of the pilot study. The feedback form completed by the teacher revealed that there were no problems during the administration and the students enjoyed the assessments. The teacher explained that the assessments took place during the last two hours of the school day and the students were tired. However, the students yet had enough time to complete the assessments, which took them approximately 15 minutes. Also the teachers said that the students "understood most instructions. They had difficulty in understanding the sentence 'Take for granted that what is said in the box is true and try to reach the correct conclusion"'. This was the first time they had to do an activity of this kind and they kept thinking of alternatives to the scenario or imagining subtext, which altered the 'take for granted' instruction.

The teacher also referred to the instructions for the administrator: 'I thought that the instructions to the administrator were too long and sometimes unnecessarily complicated'. This was a particularly interesting feedback, because it would not be considered that the problematic part could be the administration instruction. Initially, I chose to have more complicated instructions, but reassure that the process in all the

schools will be standardised. The aim of the analytical instructions was the exclusion of any potential ambiguity and vagueness in the process because it is crucial for all the schools to follow accurately the same process. Nevertheless, the feedback sent was extremely significant. Teachers are usually busy, and it is crucial to provide them with simple instructions to follow. For this reason, I decided to get further feedback for the language and the possible wordiness of the instructions. I asked two external people to judge the guides and I rephrased the instructions to make them simpler.

9.7. Chapter Summary

To sum up, based on the pilot there were slight changes made at the measurement tools. Based on the functions that Cohen, Manion and Morrison (2007, p. 341-342) suggested, this pilot study achieved its goals. Particularly:

- It ensured that the questions were suitable for the students' experience. For this reason, the measurement tools were piloted in a group similar with the targeted group.
- It led to the practicing of the coding system of data analysis. As an additional benefit of the pilot study was the improvement of the grading system for the second creativity activity.
- It provided data for distractor analysis. For each item, I checked how many students answered each of the three options A, B and C and therefore I considered whether there is a specific distractor who confused the students, because it was too believable or tricky.
- It suggested that some items might have low discrimination.
- It confirmed the clarity of the instructions and the items. Comments were provided by the students and the teacher and led to the reduction of the vagueness or difficulties in wording. In order to reassure this, I asked the teacher to right on the answer sheet the wording for which the students asked clarification. Furthermore, the students provided their own feedback written on the forms. There was just one phrase that was judged problematic 'taken for granted' and it was decided to be replaced by 'what you read is definitely true'.
- It gave the opportunity to receive comments on the type of questions and its format. The students did not have problems with the format of the questions. However, the lack of a thinking cloud on the last page made a student wonder and therefore I

decided to also include a cloud also in the last page. Unexpectedly, in this category I had the feedback regarding the administration guide. The instructions were judged as too complicated. This led me to the decision to redesign the administration guide by simplifying it but ensuring the standardised process.

- It provided a realistic image for the appropriateness of the questions. There was no missing data in this case. However, the missing data, as it has been explained before, could have been revealing concerning the item difficulty. Furthermore, based on the teacher comments the questions were appropriate for the majority of the students. However, the teacher clarified that the SEN students found difficult to reply to the complicated thinking problems.
- It highlighted omissions or irrelevant information in the forms. The students' comments revealed the omissions in some problems and there were phrases that were amended.
- It provided feedback on the attractiveness and appearance of the questionnaire. I received positive feedback by the teacher regarding the reaction of the students for the appearance of the questionnaires. For this reason, I decided not to make any changes in the appearance.
- It revealed how much time the questionnaire requires to be completed and whether it is too extensive or too short. I decided not to be present in order to pilot it in the exact same way as it will be implemented in the actual study. Therefore, the classroom teacher gave me feedback on the time which is needed for the forms to be completed. The allocated time was 30 minutes. However, the students needed 10-20 minutes to complete it. Additionally, the students commented that the time was too long. When the questionnaire needed less time than expected to be completed, there were two possible options. The one would be to add some questions in the questionnaire. The other was to change the suggested time from 30 to 20 minutes. Concerning the first option, I included a second problem-solving question. I recognized that the ability of the students to solve problems might be context-dependent. Thus, I included a second problem to enable the students to demonstrate their problem-solving skill in two different contexts. The addition of more items was not justified when the utility of the assessment was concerned. Secondly, the test demands already by the students to think. I thought that adding more thinking problems will be unreasonably demanding by the students. I rejected the second

option, because I thought that maybe students in a different cohort might work on a different pace. It would be better for teachers and students to have the pleasant surprise of finishing earlier, rather than having students who are rushed to finish because of the limited time. Moreover, the specific measurement tool does not aim to measure the thinking speed of the students and hence the time is not a factor.

10. Results of the Systematic Literature Review: P4C impact on cognitive and non-cognitive factors

The first research question of this thesis investigates the existing evidence concerning the effectiveness of the P4C programme. Lipman (2003) argued that students develop their critical, creative and caring thinking by taking part in a Community of Enquiry. This systematic literature review scrutinised whether these claims are indeed real and discussed the skills that P4C improves according to the published evidence. This chapter also shows the research literature gaps concerning the effectiveness of the programme. The review focused on particular characteristic of the studies:

- the research design of the studies
- the country where the studies were conducted
- the cognitive or non-cognitive skills that P4C could have an impact on
- the intervention and its length
- the follow-up of the participants after the end of the intervention
- the characteristics of the participants with main focus on age and gender
- the sample size of the intervention and comparison group
- the sample attrition (dropout) from pre-test to post-test
- the pre-test equivalence (or lack of equivalence) between the performance of the intervention and the comparison group before the implementation of the intervention
- the post-test results
- the reported means and standard deviations to calculate the effect sizes and enable the comparison of the findings coming from different studies

Originally, the effect sizes were calculated based only on the post-test performance of the two groups (Ventista, 2018b). However, there were a few studies where the effect sizes based on the post-test gave an inaccurate image of the programme effectiveness because of initial imbalance in the performance of the two groups in the pre-test. Therefore, in this revised version I consider also the pre-test performance for the calculation of the effect sizes.

This calculation of effect sizes was impossible for one study. Reznitskaya et al. (2012) reported pre-test equivalence. However, they used different measurement tools in the pre-test and post-test. In order to calculate the effect sizes by considering both

the pre-test and post-test performance, the scores of the tools should have been turned into z-scores because they were reported in different scales. This was not possible because I did not have access to the raw data. Even in that case it might have been unfair to calculate the effect sizes in this way, because the two tools measured different skills. Pre-test measured the performance of the two groups in reading comprehension and in a persuasive essay and established the equivalence between the two groups. However, post-test measurement tools examined the transfer of argumentation development, student questioning and the skills of elaborated description. Thus, for this study the effect sizes are still based only on the performance of the two groups during post-test assuming that there was equivalence of the two groups in the post-test.

The effect sizes of Tian & Liao (2016) should not be considered directly comparable to the others because the study reported only paired standard deviations, Therefore, there was a compromise in the calculation of the effect sizes.

Before reporting the results, it is necessary to discuss two studies whose calculated effect sizes are considered particularly untrustworthy. Firstly, Nia (2014a) conducted a study to investigate P4C impact on the anger of teenagers. The study reported sufficient information for the effect sizes to be calculated. However, every occurrence of reporting means and standard deviations in the paper for both pre-test and post-test measurement for both groups suggests equal values for the mean and equivalent standard deviation (see Table 7 in the appendix 5a). This is unlikely and suggested a typographical error.

Similarly, Nia (2014b) published a second article - probably the same study with the same sample - about the P4C impact on different types of self-esteem. The reporting appears normal compared to the previous study since means and standard deviations do not appear to be identical. However, the reporting of the public self-esteem domain is odd. Whilst the mean score of the comparison group increased from 20.4 (pre-test) to 22.03 (post-test), the mean score of the intervention group decreased from 47.19 (pre-test) to 2.18 (post-test). This is a huge decrease in the performance of the intervention group. It is probable that there was a typographical error in the reporting of the mean of this group.

10.1. Research Design

The main inclusion criterion for the studies in the review was their research design and only experimental, quasi-experimental and studies with a suitable comparison group

were examined. The existence of a comparison group and pre and post-test measurements were judged necessary indicators of the quality of causal studies. Thirty nine studies were included in the review. However, there were only a few studies considered to have a strong research design, including a recent randomised controlled trial with randomisation at school level in England (Gorard, Siddiqui & See, 2015) and a randomised controlled trial with more than 100 participants in each of the two groups (Reznitskaya et al., 2012). Some of the small-scale studies reported randomisation of participants within the groups (Hedayati & Ghaedi, 2009; Lam, 2012; Marashi, 2008; Sanz de Acedo Lizarraga et al., 2003; Topping & Trickey, 2007; Trickey & Topping, 2006). However, these studies are not better in quality compared to the other retrieved small-scale studies with a comparison group, because the number of participants was relatively small (≤ 100 for the smallest group). When the number of the participating units in a study is very small, then the randomisation is more likely to lead to concealed imbalance (Gorard, 2013, p.128). Random allocation of participants between the groups is not the only quality indicator discussed by this review. The quality of the studies retrieved is discussed later in this chapter (section 10.10).

10.2. Location of the study

Although the first P4C studies were conducted mainly in the USA and UK, currently there is research evidence from other countries. This is indicative of both the interest of the research community and its popularity in schools across the world. The programme is currently practiced in approximately 60 countries (SAPERE, 2015a).

10.3. Targeted skills

The impact on a range of skills was addressed. There were studies which examined the impact of the programme on cognitive and non-cognitive skills. Some interventions examined the impact on attainment, whilst others examined the impact on non-cognitive skills, such as social skills. Some of the studies focused on specific aspects of attainment, such as whether P4C can support students learning English as a foreign language (Tian & Liao, 2016). Studies also examined the impact on psychosomatic disorders (Shatalebi & Hedayati, 2016) and anxiety (Tian & Liao, 2016).

There is a study which examined the impact of the programme on moral judgment (Jahani, Nodehi & Akbari, 2016) and one on moral autonomy (Schleifer et al., 2003). None of the studies provided sufficient reporting to measure the impact on

moral judgment. This might imply that it is difficult to measure moral judgment. Ioannou, Chatziefraimidou and Ventista (under review) argued that moral education can be linked with different theories of ethics and different teachers perceive this type of education differently. Hence, it might be difficult for P4C studies to claim that they measure the effect on moral judgment, because this would suggest that there is an intented moral judgment to be achieved. There is no agreement on a desired moral judgement or outcomes of moral education.

Most of the interventions investigated the impact of the programme on reasoning skills (Cooke, 2015; Fair et al., 2015a, 2015b; Fields, 1995; Gorard, Siddiqui & See, 2015; Jenkins, 1986; Lam, 2012; Marashi, 2008; Reznitskaya et al., 2012; Säre, Luik, & Tulviste, 2016; Sasseville, 1994; Slade, 1989; Sprod, 1998) and a study examined the impact of the study on critical thinking dispositions (Rahdar, Pourghaz & Marziyeh, 2018). This is not surprising since Lipman, who is the father of P4C, argued that P4C fosters critical thinking (Lipman, 2003). What might be surprising is that even though he also argued that P4C fosters creativity (Lipman, 2003), only a few studies examined its impact on creativity (Abadi & Akbari, 2017; Jahani & Akbari, 2016; Pourtaghi, Hosseini & Hejazi, 2014).

This might be due to the fact that critical thinking can be more easily operationalised in reasoning skill items, whilst creativity can be considered to be a broader concept requiring assessments with open-ended items and subjective marking. Another reason which could explain this finding is the difficulty in developing creativity assessments. Predominantly, the P4C research tradition is associated with the New Jersey Test of Reasoning Skills (Shipman, 1983). This test was created by Virginia Shipman who worked in the Institute for the Advancement of Philosophy for Children (IAPC) at Montclair University, where Matthew Lipman also worked. The test included 50 items evaluating general, hypothetical and causal reasoning, assuming, induction, good reasons, syllogisms, contradiction, standardisation and conversion (Morante & Ulesky, 1984).

Lipman and the first P4C adherents did not develop an instrument to measure creativity. Consequently, it is not surprising that most P4C studies scrutinise its impact on reasoning instead of creativity since it appears to be more guidance on how to evaluate these skills. The researchers who choose to evaluate creativity have to decide upon an existing tool from a different field or construct a new one for creativity

assessment. For example, in one of the retrieved studies the researchers used the Torrance Test to measure creativity (Pourtaghi, Hosseini & Hejazi, 2014).

10.4. The intervention and its length

In most of the retrieved studies, the intervention group received only P4C. However, Sanz de Acedo Lizarraga et al. (2003) examined the combination of P4C with the Instrument Enrichment Programme and Project Intelligence. The intervention received by the participants in this study was called *Portfolio* and, therefore, the impact and the effect sizes cannot be attributed solely to P4C. Even though in all the other studies the participants received P4C intervention, the programme implementation varied between them.

Considering the length of the intervention, it usually lasted an academic year or less (See Table 7 in the appendix 5a). The intervention was sometimes too short for meaningful results to appear in the assessment results.

10.5. Follow-up study

Only a few studies (Colom et al., 2014; Fair et al., 2015b; Sanz de Acedo Lizarraga et al., 2003; Topping & Trickey, 2007; Youssef, Campbell & Tangen, 2016) incorporated follow-up in their design. Without an adequate number of longitudinal studies, there is no strong evidence of the long-term impact and the retention of the effects of the programme.

Colom et al. (2014) claimed to follow the same participants for twelve school years (from six to eighteen years old). However, the reporting of the study was inadequate for effect sizes to be calculated and the effectiveness of the programme to be discussed.

10.6. Participants

Even though P4C, as the name suggests, focuses on children, this systematic literature review identified studies with participants from a wide age range. There were studies with participants in kindergarten and students younger than six years old (Giménez-Dasí, Quintanilla & Daniel, 2013; Jo, 2001; Säre, Luik, & Tulviste, 2016; Schleifer et al., 2003) and studies with participants older than twelve who can be considered teenagers (Abaspour, Nowrosi & Latifi, 2015; Lam, 2012; Pourtaghi, Hosseini & Hejazi, 2014; Sanz de Acedo Lizarraga et al., 2003). However, the few studies investigating P4C impact on teenagers usually involved students in the early phase of

adolescence. There was only one study which examined the impact on P4C on teenagers aged 16-17 years old (Tian & Liao, 2016).

The participants of one study were university students (Abadi & Akbari, 2017). In fact community of enquiry has been increasingly popular in universities. Demissie (2017) discussed how useful it was to implement P4C with teacher educators. P4C can help adults, not only children. For example, in the case of teacher educators, they can develop their reflective thinking, which is necessary for their teaching practice. However, it could be argued that P4C refers to philosophy in childhood, and teaching philosophy during adolescence and adulthood is a different programme. Community of enquiry can be implemented with different age groups, but the term P4C should not be used for different age groups.

In addition, the examination of the particular characteristics of participants was judged to be crucial. This review presents the country that the research was conducted, the age and the gender of the participants (see Table 7 in the appendix 5a), but there are not the only characteristics which could influence the research results. Information such as the socioeconomic background and the type of education that the school provides can add further information which explains the result of the study better. For example, a study in Spain (Giménez-Dasí, Quintanilla & Daniel, 2013) had pre-school students who were Caucasian from middle-class families attending a private, non-religious school near Madrid as participants. In another example, participants in a Hong Kong study came predominantly (90%) from middle or working-class families (Lam, 2012).

Despite recognising the importance of the special characteristics of the participants in each study, it would probably be overly complex to consider all these elements at once combined with the interpretation of P4C impact. There are no claims concerning generalisation on or representation of a particular type of population. Hence, no attempt to generalise the results will be made because there is no clear statement of the population represented by the participants of each study.

There were some studies in which all the participants had the same sex (Abaspour, Nowrosi & Latifi, 2015; Pourtaghi, Hosseini & Hejazi, 2014; Shatalebi & Hedayati, 2016) conducted in Iran. The option of involving only single-sex participants does not seem to be grounded upon a justified research decision or linked to the research questions. Instead, it probably derives from the single-sex education in the country. Shatalebi and Hedayati (2016) mentioned in their title and abstract that their

study involved only boys aged 9-11 years old. However, in the section of the sample in the article they mentioned 'population consists of all female students' (p.4).

Slade's study (1989) had only female participants. This decision was grounded on the research question. Female students were considered weaker at mathematics, and the study aimed to investigate whether P4C potentially improve their mathematical ability.

10.7. Sample size

Concerning the number of participants, there were only seven studies with more than 200 participants in both groups (Colom et al., 2014; Fair et al., 2015a; Gorard, Siddiqui & See, 2015; Reznitskaya et al., 2012; Sasseville, 1994; Siddiqui, Gorard & See, 2017; Youssef, Campbell & Tangen, 2016). Probably the large-scale studies provide more trustworthy results. Most of the studies have a small sample size and this can be considered a weakness in their research design. This is in line with the finding of the meta-analysis conducted by García-Moriyón et al. (2004). The authors also found that studies used small samples and they expressed a consideration about generalising results occurring by studies with small sample sizes.

Concerning the sample, this review did not consider the numerical balance between the comparison and the intervention group as an indicator of the project quality. Gorard (2013, p.128) argued that the two groups do not have to be arithmetically equal, but he suggested a limit with the one group being up to three times bigger than the other, with the comparison group usually being bigger as it increases the power with low research cost. However, the comparison group and the intervention groups were equal or almost equal concerning the number of participants in most of the studies (see Table 7 in the appendix 5a). There were a few studies in which the intervention group was bigger than the comparison group (Colom et al., 2014; Fair et al., 2015a, 2015b; Sasseville, 1994; Topping & Trickey, 2007; Trickey & Topping, 2006) and there were only two studies with a bigger comparison group than intervention group (Lam, 2012; Siddiqui, Gorard & See, 2017).

10.8. Attrition

Research attrition is a central indicator of the trustworthiness of the findings. Nevertheless, many research studies did not report attrition (Table 7 in the appendix 5a). Fair et al. (2015a) did not report attrition between the pre-test and the post-test, but they did report the attrition for the follow-up cohort of the 7th graders (Fair et al.,

2015b). Some studies implied that the sample was retained in the tables which report their findings since they stated the same number of participants (N) in the pre-test and the post-test. The studies that retained their sample were small-scale and involved short-term P4C intervention. This finding is in line with what Gorard (2015) argued. The larger the study and the longer it lasts, the more attrition it is likely to have. Therefore, it is not surprising that many small-scale studies retained their sample (Abaspour, Nowrosi & Latifi, 2015; Jenkins, 1986; Jo, 2001; Pourtaghi, Hosseini & Hejazi, 2014; Sanz de Acedo Lizarraga et al., 2003; Slade, 1989; Sprod, 1998; Tian & Liao, 2016; Williams, 1993).

10.9. Pre-test equivalence

Examination of the baseline assessment was judged necessary. This is why this chapter is a revision of what was presented by Ventista (2018b) and considers the pre-test scores. Some studies had initial group imbalance in their performance at the pre-test. For instance, Pourtaghi, Hosseini and Hejazi (2014) and Lam (2012) conducted small-scale trials with the comparison group performing better than the intervention group at the pre-test. On the other hand, studies such as Fair et al. (2015a, 2015b) reported that the intervention group was performing better than the control group in the pre-test.

10.10. Impact of the programme on cognitive and non-cognitive skills

The design of the studies was used to evaluate the quality of the studies. Table 10.1 summarised the quality of the 39 studies included in the systematic literature review. All of the studies have a comparison group, but most of the studies are low-quality based on their design.

Table 10.1. Quality of Research Design and Reporting of the Studies included in the Systematic Literature Review

Quality Indicators	Number of Studies
★★★★★	2
★★★★	1
★★★	8
★★	15
★	7
0	6
Total Number of Studies	39

The quality of the studies demonstrates that the evidence should be interpreted with cautiousness. The sample size of the studies does not allow generalization to

populations. However, the studies provided indicators about the programme effectiveness. Effect sizes for the studies were calculated in order to create comparable results and investigate whether and on what skills P4C has an impact. Studies which did not report sample size, means and standard deviations could not have had their effect sizes calculated (Table 7 in the appendix 5a). As mentioned earlier in this chapter, effect sizes of two of the studies (Reznitskaya et al., 2012; Tian & Liao, 2016) were calculated with some compromises and therefore they should not be considered directly comparable with the other effect sizes.

Furthermore, when the effect sizes are under consideration, the domain of potential improvement of the study should be evaluated. For example, one of the studies (Abaspour, Nowrosi & Latifi, 2015) appeared to have generally negative strong effect sizes in the areas of disappointment and instability. This actually means a big positive impact of P4C on a negative domain. In other words, if it is a negative scale a negative effect size is treated as positive.

According to Cohen (1988), an effect size is considered big when $d \geq 0.80$ and medium when $d = 0.50$. Based on his recommendations, this review accepted for $d < 0.50$, the effect size was small. When $0.50 \leq d < 0.80$, the effect size was considered medium. Finally, for $d \geq 0.80$ the effect size was considered big.

In the appendix there is a table (Table 7) which presents all the information about the studies and the effect sizes calculated. It becomes apparent that samples with small sample might receive two stars in the evaluation because they retain their sample size (such as Abadi & Akbari, 2017; Rahdar, Pourghaz & Marziyeh, 2018; Shatalebi & Hedayati; 2016). It also becomes apparent that studies with small sample tend to retain their sample, whilst studies with big samples report attrition. Table 10.2 presents the relationship between of the quality of the studies and the effect sizes they reported. Some studies report more than one effect sizes and this is why the number of effect sizes does not match the number of the studies presented in the Table 10.1.

Table 10.2. Quality of Studies in relation to reported Effect Sizes (P4C impact)

Number of Stars	Big	Medium	Small	Negative Impact
5 stars	1	1	4	1
4 stars	0	0	2	6
3 stars	6	2	3	4
2 stars	8	5	5	10
1 star	1	0	0	1

0	6	0	1	1
Total	22	8	15	23

Studies of high-quality usually reported small positive impact whilst studies of lower quality (1 and 0 stars) big. Big effect sizes can be attributed to research design factors and not the P4C intervention itself. All of the studies of low quality had particularly small samples and it is common for studies with small samples often report bigger effect sizes (Gorard & Gorard, 2016, p.483; Slavin & Smith, 2008). Thus, the design is likely to have been the cause of the observed big effect sizes rather than the actual P4C effectiveness.

Finally, table 10.3. presents the effect sizes reported in the studies in relation to the skills examined by these. For table 10.3 it has to be noted that some studies examine more than one skill. This is why there is no agreement between the number of studies presented in Table 10.1 and the effect sizes in Table 10.3. Table 10.1 refers to the actual number of studies, whilst tables 10.3 refers to the reported effect sizes, with some studies having reported more than one effect size.

Table 10.3. Skills being examined in the retrieved studies in relation to calculated effect sizes

	Big Positive	Medium Positive	Small Positive	Negative	Total
Reasoning	5	3	5	0	13
Questioning	0	1	0	0	1
Critical Thinking (Dispositions)	1	0	0	0	1
Creativity	5	0	0	0	5
Self-esteem, self-efficacy, confidence	2	1	1	4	8
Social skills, Co-operation, Talkativeness	2	0	2	5	9
Well-being	0	0	0	3	3
Literacy (reading, meaning construction, comprehension, writing)	0	2	4	3	9
Disorders, Anxiety, Anger	0	0	1	2	3
Maths	0	0	1	1	2
Non-cognitive skills (e.g. emotion comprehension, motivation)	7	1	1	5	14

10.10.1. Critical Thinking and Reasoning Skills

Lipman (2003) supported that students develop their critical thinking by taking part in a Community of Enquiry. This is one of the main acceptances amongst P4C adherents. For example, SAPERE (2015c) mentions that P4C develops the 4 C's for students, one of which is critical thinking. There may be studies which examined the impact of the programme on critical thinking skills. For example, Karadağ and Demirtaş (2018) examined the impact of the programme on 5 and 6-year-old students' critical thinking skills in two classrooms. The one was in a private school and the other in a state school. However, both groups received the intervention. This study had no comparison group. There is no single study with a comparison group which examined the impact of the programme on critical thinking skills. There is only one recent study published which examined the impact of the programme on critical openness and reflective scepticism (Rahdar, Pourghaz & Marziyeh, 2018). Earlier in this thesis, it was discussed that somebody might value critical thinking without having critical thinking skills. Also, dispositions are difficult to measure. Except for one study examining critical thinking dispositions, the other studies examined reasoning skills.

Therefore, it cannot be claimed that P4C develops students' critical thinking. However, it can be confidently said that P4C develops the reasoning of students. Based on the published evidence and their 13 calculated effect sizes, which report the impact of the programme on reasoning skills, there is no negative impact of the programme reported. The size of the impact might vary, but the performance of students in reasoning skills always improve after their participation in P4C sessions.

Furthermore, two of the studies report long-term impact of P4C on reasoning skills (Fair et al., 2015b; Topping & Trickey, 2007). The one was a follow-up study (Fair et al., 2015b), whilst the other (Topping & Trickey, 2007) examined the impact of the programme after two years of implementation. Both studies found that students in the intervention group developed their reasoning skills more than the students in the comparison group.

10.10.2. Questioning

Lipman (2009) emphasised the primacy of questioning in P4C dialogue. Ventista and Paparoussi (2016) also argued that in the Community of Enquiry the prominence of questioning compared to answering. However, only one study examined the impact of the programme on the questioning (Reznitskaya et al., 2012). This study was graded with 5 stars for its research design and found medium positive effect size on students'

questioning. This is a positive indicator about the effectiveness of the programme on this area, but more evidence is needed to establish a causal relationship between increased questioning and P4C implementation.

10.10.3. Creativity

The evidence concerning the impact on creativity is limited (Abadi & Akbari, 2017; Jahani & Akbari, 2016; Pourtaghi, Hosseini & Hejazi, 2014). For one of these studies the effect size was not possible to be calculated because of insufficient reporting (Jahani & Akbari, 2016).

The calculated effect sizes for the impact of the programme on creativity were positive and big. Both studies were conducted in Iran. Abadi and Akbari (2017) examined the impact on university students, whilst Pourtaghi et al. (2014) examined the impact on secondary-school boys. Both studies received two stars in the grading of the quality of the studies because they retained their sample. The one study had only 30 students in each group (Abadi & Akbari, 2017) and the other only 16 students in each group (Pourtaghi, Hosseini & Hejazi, 2014). Therefore, their sample was particularly small to lead to any trustworthy and generalisable results about the impact of the programme. No study examined the impact on primary school students.

10.10.4. Self-esteem

P4C gives the opportunity to the pupils to freely express their opinion. There were a few studies which examined the impact of the programme on self-esteem, self-efficacy or confidence. These studies showed mixed findings about the impact of the programme on self-esteem. Nia (2014b) examined the impact on four different types of self-esteem and reported that students' self-esteem in relation to education and family increased. However, as it has already been mentioned, there was probably a typo in the reporting of 'public self-esteem' and particularly the mean for the intervention group (see Table 7).

Siddiqui, Gorard and See (2017) found small positive effect size for students' self-confidence. This study was graded with four studies and its results are trustworthy. However, both groups reduced their self-confidence in the post-test. The decrease of self-confidence in the intervention group was smaller than the comparison group and the calculated effect size is small and positive. The age might also play a role in the decrease of the self-esteem scores.

Similarly, a study graded with three stars of quality (Youssef, Campbell & Tangen, 2016) reported negative self-esteem for the intervention group. The mean score of the students in the comparison group increased in the post-test, whilst the mean score of the students in the P4C group dropped.

It might be questioned why the self-esteem of the students does not increase after P4C implementation. It would be expected that the students would be more confident to express their opinion after practicing this skill. However, P4C supports students to recognise ambiguity in language. Students recognise that there is not always a right answer. Decrease in the scores of self-confidence might show that students are less stringent when their points of view are concerned and they accept the idea of being wrong. It might be the realisation that they might be ignorant or only knowledgeable on a topic, which makes them less confident.

It has to be mentioned that Year 4 and Year 5 pupils from the intervention group in an unpublished evaluation conducted by Swain, Cara and Litster (2014) reported that they felt that their reasoning, thinking, reading, listening and writing skills improved. Therefore, in some cases P4C might increase students' confidence and self-esteem.

10.10.5. Social Skills

A few studies report the effect of the programme on the interaction with classmates. P4C aims to transform the classroom into a Community of Enquiry and therefore it would not be surprising that this skill was examined as a potential impact of the programme. Social skills are easier observable compared to other skills, such as self-esteem. Therefore, studies might aim to measure them. The findings of the impact of the programme on social skills are inconsistent. However, they should be examined in relation to the quality of the studies.

The study graded with four stars (Siddiqui, Gorard & See, 2017) examined the impact of the programme on communication skills, sociability and teamwork. The P4C group performed poorer in these areas compared to the matched group. However, the study graded with three stars (Youssef, Campbell & Tangen, 2016) found reported a small positive effect size in the pro-social behaviour of the intervention group. Pro-social behaviour refers to behaviour which aims to help the others. As Lipman (2003) stated P4C can help the caring thinking of students. The findings of this study indicate that this might be true.

Säre, Luik, and Tulviste (2016) found a big positive effect size on talkativeness of students. Therefore, the interaction with their classmates might increase with P4C. If we accept social constructivism and that pupils learn by interacting with their peers, this might be very important. Also, 4-year old students and 5-year old students increased their knowledge about interacting with classmates after participating in P4C sessions (Giménez-Dasí, Quintanilla & Daniel, 2013). Students' voice is central to P4C sessions and students seem to interact more with their classmates in this dialogic context. However, this does not mean that students express that they can work better in groups and there is no evidence that their co-operative skills increase.

10.10.6. Well-being

Consistent evidence exists concerning the impact of the programme on the well-being of the students after their participation in P4C programme. High-quality studies support that the well-being of the students decreases after their participation in P4C session (Siddiqui, Gorard & See, 2017; Youssef, Campbell & Tangen, 2016). However, it has to be noted that the mean scores are based on self-reported questionnaires. This might be an indicator that P4C raises the awareness of students and therefore they can easier identify threats to their well-being compared to students in the comparison group. The studies report students' perceived well-being and not their well-being.

10.10.7. Cognitive skills

Concerning their cognitive skills and particularly their attainment, Table 10.3 reports their attainment related to their literacy separately to their attainment related to Maths. P4C is a dialogic intervention, which encourages students to question, use abstract concepts and express their opinion. This intervention does not seem related to the Maths ability of students. Therefore, the table reports these two separately. There are mixed results concerning the impact of the programme on attainment. Some of them support a positive impact on attainment and cognitive skills and some negative. There are more studies to suggest that P4C can have a positive impact on the literacy skills of students, concerning reading comprehension and writing.

Specifically, Gorard, Siddiqui and See (2015) found a small positive impact of the programme on students' reading and writing. This is reported separately because this study was graded with five stars for its research design. Therefore, its results are trustworthy. Similarly, Youssef, Campbell and Tangen (2016) reported small positive

impact of the programme on reading comprehension. This study was graded with three stars. Consequently, high-quality studies report small positive impact of the programme on literacy.

Concerning Maths, Youssef, Campbell and Tangen (2016) reported negative impact of the programme on the interest of the students in Maths. Since P4C turns the attention of students in philosophical discussions, it is likely to decrease their interest in different domains. Gorard, Siddiqui and See (2015) reported a small positive impact of the programme on the maths ability. This finding is difficult to be explained because P4C does not seem to have any direct association with Maths.

10.10.8. Disorders

Studies reported the impact of the programme on disorders, anxiety and anger. There are two negative effect sizes reported in this category. This two negative effect sizes though imply positive results for P4C, because it is good that P4C reduces psychosomatic disorders (Shatalebi & Hedayati, 2016) and anger (Nia, 2014a). However, Tian and Liao (2016) reported that P4C increases English learning anxiety when English is taught as a foreign language. It is possible that P4C engages the students in a dialogue which is challenging even in their native language. It requires from the students to use abstract concepts and create arguments. This might be difficult in a foreign language.

10.10.9. Non-cognitive skills

There are many studies which examine the impact of the programme on various non-cognitive skills. This might be due to the fact that the programme is widely believed to improve thinking and have wider outcomes more than attainment. P4C indeed increased the non-cognitive skills of students and this is a consistent finding coming from different studies. Despite the fact that in the table 10.3 five of the effect sizes in the category non-cognitive skills appear to be negative, three of them imply positive findings. This is because they refer to disappointment, grandiosity and instability (Abaspour, Nowrosi & Latifi, 2015). It is positive that P4C decreased these three traits in the students who attended the programme.

Abaspour, Nowrosi and Latifi (2015) also reported a negative effect size for the impression management after the students attended P4C sessions. This finding can be interpreted in both ways. This might mean that P4C has a negative impact on a non-cognitive trait. However, it might also be interpreted positively. Students in P4C group

might be more honest and less caring about managing their expressions or 'manipulating' the opinion of others.

Big effect sizes were reported for English learning motivation (Tian & Liao, 2016), adaptability and meta-cognition (Cooke, 2015). Moreover, cogency is included in this category with and in the reporting of high effect sizes (Cooke, 2015). There can be an argument that this could also be included in the effect sizes of reasoning. However, I categorised it here because I think it is not simply linked to reasoning, but it is a broader skill than this.

There is no consistent evidence about the impact of the programme on emotion comprehension. Giménez-Dasí, Quintanilla and Daniel (2013) reported a big positive effect size when 5-year-old students participated in P4C sessions, but negative impact when 4-year-old participated in similar sessions. It has to be noted that studies graded low, such as Cooke (2015) with one star and Giménez-Dasí, Quintanilla and Daniel (2013) with zero stars, reported big effect sizes, whilst the study graded with four stars (Siddiqui, Gorard & See, 2017) reported a very small positive effect size in relation to determination.

Even though a conclusion about the magnitude of impact cannot be reached, it seems that P4C develops some non-cognitive skills of students, such as determination. This is a consistent finding between various studies. It might be questionable how this happens when many of these studies are only short-term and it could be assumed that these skills might need more time to change. Consequently, it can be questioned whether these skills are in fact malleable or it is due to some research design flaws.

10.11. Discussion

This chapter discussed the quality and the results of 39 studies evaluating the effectiveness of P4C. Only a few studies were large-scale. The attrition and the missing data were not always reported. Hence, the quality of some studies was questioned. Overall, there were only a few well-designed studies (see Table 10.1.). Therefore, there is still much room for research to shed light on stronger evidence about the effectiveness of the programme.

There are various literature gaps to be covered in P4C research. Concerning the domains that the programme is expected to have impact, Lipman (2003) supported that by taking part in a Community of Enquiry students develop critical, creative and caring thinking. Concerning these specific skills, there is sufficient evidence to support the

effectiveness of the programme on reasoning skills (Table 10.3.). There were only a few studies which were poorly designed examined the P4C impact on creativity (Jahani & Akbari, 2016; Pourtaghi, Hosseini & Hejazi, 2014). Therefore, Lipman's (2003) claim is still unproven. No study strictly defined caring thinking to examine P4C impact on it. There was positive evidence regarding the impact of the programme on non-cognitive skills.

Lipman (1985) suggested that proficiency in elementary reasoning skills is associated with school performance. By saying this he meant that reasoning skills are a pre-requisite to academic success. They do not ensure success, but they are necessarily required for the success to be achieved. This entails that Lipman believed that there is a link between attainment and reasoning skills and since P4C develops the latter, then positive impact could be expected on the former. However, it is known that P4C does not directly teach linguistics or mathematics. Some of the studies retrieved focused on the P4C impact on attainment. There is limited and contradictory evidence about the effectiveness of the programme on this area (Table 10.3.).

Based on the current studies, the programme does not appear to have any detrimental impact on any cognitive or non-cognitive domain. On the contrary, P4C improves some cognitive and non-cognitive skills. P4C has a positive impact on reasoning skills which is also retained for years after the end of P4C implementation as follow-up studies have demonstrated. This is in line with findings of a recently completed meta-analysis conducted by Yan (2017). This meta-analysis also reported big positive impact of P4C on reasoning skills. Consequently, P4C should be implemented at primary schools.

Following this review, further investigation of many areas is recommended. This thesis attempts to discuss some these areas in the following chapters. Evidence is needed regarding the impact of the programme on thinking skills. This will be discussed in the next chapter with the results of the comparative evaluation study. As in the case of reasoning skills, there might be other skills on which the programme might demonstrate an impact after the end of the intervention or after a long-term implementation. Chapter 12 examines the long-term impact of the programme on attainment. Providing robust research evidence can contribute towards the wider acceptance of the programme by teachers and head-teachers, support evidence-based policy and lead to the P4C introduction in the school curriculum in a more systematic way.

11. Results of the Comparative Evaluation Study: The impact on Thinking Skills

This thesis examined whether P4C has an impact on pupils' critical thinking and creativity. P4C adherents discuss the impact of the programme on these skills. However, the systematic literature review showed that there was no single study with a comparison group which examined the impact on critical thinking overall. The existing evidence regarding the impact of the programme on creativity was weak. In this chapter, results of the evaluation study are presented and discussed. Regressions investigated variables which could possibly predict pupils' good performance in thinking skills.

11.1. The impact of Philosophy for Children on Critical Thinking

There is available evidence about the impact of the programme on reasoning as presented in Table 10.3. The published studies consistently found positive impact of the programme on reasoning. This thesis examined the impact on critical thinking overall. Before presenting the results, there is a presentation of some descriptive statistics. These present the way that the students responded to the thinking problems.

11.1.1. Descriptive Statistics

The difficulty of each item was discussed in the pilot study. Nevertheless, it can also be found by the number of students in both groups who responded each item right in the comparative evaluation study. Items which can discriminate the performance of the students are needed for the assessment. Table 11.1. and 11.2 present the number of correct and wrong responses for each item in the pre-test and post-test respectively. No item was responded right or wrong by all students.

Furthermore, these two tables report the number of blank and double-marked responses for pre-test and post-test. There were not many blank and double-marked responses in the pre-test and the post-test. Both of these categories were scored with zero in the multiple-choice questions because they were no right response for the questions and therefore they gave no marks. This scoring is usually adopted by Classical Test Theory. However, it is worth reporting these two types of responses,

because they are not the typically wrong responses. The number of blank responses could reveal patterns for the quality of the assessment. For example, the number of blank responses in both pre-test and post-test increased towards the end of the assessment. For example, the thinking problem 7 in the pre-test had 29 blank responses, whilst the first item had only 5.

It is likely that the order that the questions appeared on the test and not the content was the reason why this happened because there is a clear pattern of gradually increased blank responses. This confirmed Traub and Rowley (1991), who argued that in a timed assessment with strict time limits the items which appear last are more likely to be affected by the time restrictions. Therefore, the missing data might indicate that the students did not have enough time to complete the assessments. However, it might also indicate that some of the students were no longer concentrated or motivated towards the end of the assessment. Instead of guessing they chose to leave this question blank.

The number of double-marked responses is also reported separately. As in the case of blank responses, this is a category which should be examined separately. Since the assessment has only one right answer, double-marked responses might reveal that the items are not well-designed and they leave room for a second answer to be considered equally correct. Since I designed the questions, I examined this carefully to see whether there was a problem with the design. The number of double-marked responses was not high in order to reveal that a significant number of students could not identify the right answer. Therefore, I had no evidence to believe that there was a problem with the test construction.

Table 11.1. Descriptive Statistics for Thinking Problems(TP) in Pre-test

N of responses in the Critical Thinking Problems in Pre-test	**TP1**	**TP2**	**TP3**	**TP4**	**TP5**	**TP6**	**TP7**
Correct responses	297	318	632	391	540	73	468
Wrong responses	512	489	173	412	262	722	320
Blank (no responses)	5	6	12	11	13	20	29
Double-marked responses	3	4	0	3	2	2	0
Total	817	817	817	817	817	817	817

Table 11.2. Descriptive Statistics for Thinking Problems (TP) in Post-test

N of responses in the Critical Thinking Problems in Post-test	**TP1**	**TP2**	**TP3**	**TP4**	**TP5**	**TP6**	**TP7**
Correct responses	262	431	384	391	189	398	272

Wrong responses	465	296	332	326	530	319	441
Blank (no responses)	9	9	14	17	16	17	20
Double-marked responses	2	2	8	4	3	4	5
Total	738	738	738	738	738	738	738

Table 11.3 shows how the scores were distributed in pre-test and post-test. It becomes apparent that the scores were approximately normally distributed, since most of the students responded correctly to 2-5 items and only a few students received extreme scores, such as 0.

Table 11.3. Number of Students responded correctly to the Critical Thinking Problems

N Critical Thinking Problems answered correctly	Pre-test	Post-test
None (0)	18	32
1	52	73
2	153	146
3	208	179
4	227	166
5	126	99
6	32	35
All (7 problems)	1	8

11.2. Critical Thinking: Results

This section examines the impact of P4C on critical thinking skills. First, it examines the impact of the programme on critical thinking and then on each skill included in this construct.

11.2.1. Calculating the Critical Thinking Overall

According to this thesis critical thinking construct consists of a combination of skills. Specifically, it was operationalised as inference, credibility, assumption identification, reasoning and problem-solving. Hence, before examining the impact of the programme on critical thinking, the data was used to examine whether these skills were different. The performance of the students in each skill was compared to their performance in the items measuring the other skills. Low correlation was found (see Tables 11.4 and 11.5) and therefore I argue that each section of the assessment measured a different facet of the critical thinking construct.

These low correlations were expected and they are in agreement with the findings of my previous research (Ventista, 2018a) with critical thinking tools.

Students who perform well in one assessment may not perform well in another, when the latter measures different aspects of critical thinking.

Table 11.4. Correlations of students' performance in different critical thinking skills (Pre-test)

	Inference	Credibility	Assumption Identification	Problem Solving	Reasoning
Inference	1	-0.290	0.036	0.002	0.022
Credibility			0.026	0.043	0.088
Assumption Identification				0.095	0.146
Problem Solving (items 6 and 7)					0.049
Reasoning (items 3 and 4)					1

Table 11.5. Correlations of students' performance in different critical thinking skills (Post-test)

	Inference	Credibility	Assumption Identification	Problem Solving	Reasoning
Inference	1	0.086	0.061	0.121	-0.055
Credibility			0.147	-0.001	0.114
Assumption Identification				0.165	0.126
Problem Solving (items 5 and 7)					0.078
Reasoning (items 3 and 4)					1

Since each section measured a different aspect of the construct, critical thinking overall was calculated as presented in the methods section. An average score was calculated since there was no reason to assume that a specific skill was more important than the others. Current literature does not suggest that a skill is of a more importance than others. A slightly negative effect size was found for the critical thinking (Table 11.6).

In order to examine, the trustworthiness of the studies, the number of cases who dropped out was compared to the number of counterfactual cases needed to disturb the finding (NNTD). From the table 11.6, it becomes apparent that after the pre-test 52 cased dropped out from the intervention group and 27 from the comparison group. Therefore, 79 cases dropped out from the study. Given the effect size and the number of cases in the smallest group, the number of counterfactual cases was only 12.

Since 79 participants dropped out from the study and the number of counterfactual cases needed to make the effect size disappear is only 12, the results of this thesis can be considered tentative. Therefore, although the results suggest that P4C has no impact on critical thinking, this finding can be considered provisional.

Table 11.6. Impact of Philosophy for Children on Critical Thinking

Critical Thinking Assessments	Pre-test			Post-test		
	N	Mean	SD	N	Mean	SD
Intervention Group	547	0.48	0.19	495	0.45	0.21
Comparison group	270	0.47	0.20	243	0.45	0.24
Effect Size	0.05			0.00		
Effect Size (Pre and Post-test)	**- 0.05**					
NNTD	12					

Lipman (2003) argued that P4C improves critical thinking in general and did not restrict his claims on reasoning skills. This study does not provide evidence to support this claim. It is apparent that none of the groups improved their performance in critical thinking assessments. This thesis earlier argued that critical thinking is a skill which can be developed. This is not confirmed by this finding. After one academic year, none groups developed their critical thinking. It can be questioned to what extent critical thinking is malleable. Neither regular practice nor P4C currently involve explicit teaching of critical thinking. I argue that the finding in Table 11.6 demonstrates that implicit teaching of critical thinking does not improve it. If critical thinking is considered an important educational aim, it can be questionable whether schooling improves critical thinking when teachers do not teach critical thinking.

According to the results of the comparative evaluation study, P4C does not have an impact on critical thinking. This could have various explanations. A possible explanation might be that the implementation in the English schools nowadays is different from the implementation that Lipman suggested. It has already been explained that even though SAPERE follows Lipman's model, there are adjustments in the implementation in the UK. For example, the introductory stimulus for the dialogue varies in the English schools, while Lipman suggested a strict and specific curriculum with novels written specifically for this reason. Furthermore, Lipman novels present a dialogue among various characters which argue and counterargue. These characters model different forms of thinking for the pupils to imitate in the classroom. This might

be one of the possible reasons explaining why the impact of the programme on critical thinking might be reduced.

Nowadays, the programme might not include systematic teaching of critical thinking as Lipman would support. P4C actually did not involve all the aspects of critical thinking measured in this test. Even if the initial curriculum suggested by Lipman included practice in all of these skills, P4C today is implemented in a more flexible way. For example, the development of skills of assumption identification and examining the credibility of sources are not usually explicitly reported and examined in P4C sessions. The emphasis is usually on reasoning and justification of opinions (Ventista & Paparoussi, 2016). Gorard, See and Morris (2016, p.161) argued that the programme aims to develop pupils' abilities of reasoning, disposition to question, argumentation and communication. The authors referred to reasoning and argumentation, but they did not explicitly refer to other skills included in the working definition of critical thinking by this thesis. Therefore, the P4C discourse focuses mainly on reasoning. By visiting the schools that took part in the study, I did not observe explicit teaching of aspects of critical thinking during the P4C sessions.

Although the finding in Table 11.6 suggests that implicit teaching of critical thinking is not effective, it does not suggest the opposite. There is no evidence to suggest that the Lipman's approach, which involved an explicit teaching of critical thinking, is effective.

11.2.2. Philosophy for Children impact on different Critical Thinking Skills

Having discussed the impact that P4C had on critical thinking, an examination of each skill is presented separately (Table 11.7). This is due to the fact that the correlations between the performance of the students in the skills in the pre-test and post-test were very low (see Tables 11.4 and 11.5). This means that each section measured a different skill. P4C might have had an impact on some of them, but not on others.

Table 11.7. Impact of Philosophy for Children on Critical Thinking Skills

Critical Thinking Skills		**Pre-test**			**Post-test**		
		N	**Mean**	**SD**	**N**	**Mean**	**SD**
Inference	Intervention Group	547	0.39	0.49	495	0.37	0.48
	Comparison group	270	0.32	0.47	243	0.33	0.47
	Effect Sizes	0.14			0.08		
	Effect Size (pre-test and post-test)	**-0.06**					

	NNTD	15					
Evaluating the credibility of sources	Intervention Group	547	0.39	0.49	495	0.56	0.50
	Comparison group	270	0.40	0.49	243	0.64	0.48
	Effect Sizes	-0.02			-0.16		
	Effect Size (pre-test and post-test)	**-0.14**					
	NNTD	34					
Reasoning (Deduction)	Intervention Group	547	0.62	0.35	495	0.53	0.38
	Comparison group	270	0.64	0.37	243	0.52	0.42
	Effect Sizes	-0.06			0.02		
	Effect Size (pre-test and post-test)	**0.08**					
	NNTD	19					
Assumption identification	Intervention Group	547	0.67	0.47	495	0.56	0.50
	Comparison group	270	0.64	0.48	243	0.50	0.50
	Effect Sizes	0.06			0.12		
	Effect Size (pre-test and post-test)	**0.06**					
	NNTD	15					
Problem-solving	Intervention Group	547	0.34	0.28	495	0.32	0.34
	Comparison group	270	0.32	0.27	243	0.30	0.32
	Effect Sizes	0.07			0.06		
	Effect Size (pre-test and post-test)	**0.00**					
	NNTD	0					

Concerning inference, the performance of the intervention group slightly decreased in the post-test, whilst the performance of the comparison group slightly increased. Given the measurement error of all assessments, it can be argued that both groups did not really change their performance from the beginning to the end of academic year. Furthermore, the intervention group was ahead to the comparison group in both occasions. Thus, P4C has no impact on students' inference skill.

Concerning the ability of students to evaluate the credibility of sources, the mean scores of both groups increased in the post-test. However, the students in the comparison group developed this ability more than the students in the intervention group. Therefore, students who attended P4C sessions developed their ability to evaluate the credibility of sources less than the students of the comparison group.

Both reasoning and assumption identification have positive effect sizes in table 11.8. Both groups reduced their performance in the post-test. The positive effect size suggests that the reduction of performance in the post-test was less in the intervention than the comparison group and it does not suggest that there was an improvement in the intervention group. With reference to problem-solving, both groups slightly

reduced their performance at the end of academic year. This might be due to the fact that items 6 and 7 in the pre-test were overall easier than items 5 and 7 in the post-test (see Tables 11.1. and 11.2). This might be the reason why the mean scores of both groups decreased.

It has to be mentioned that the NNTD in all occasions is smaller than the number of cases who dropped out (79 cases dropped out from the study). Thus, no strong claims can be made about the study findings regarding critical thinking skills. Concerning, the NNTD reported for the problem-solving, it is zero, because the effect size is also zero. No cases are needed to make the effect size disappear, because the effect size is already zero.

11.3. Regression for Critical Thinking Performance

Following the effect sizes, there was an attempt to create a regression because there were some elements that were not considered in the effect sizes calculation. Some of the schools were involved in P4C sessions more than one year and thus they did have different starting point. This could be considered a factor which changed the effect that the intervention schools appear to have. What if the programme has positive effect only on the first year or needs time to show some effect? The years of the participation in the programme might be an indicator. Moreover, the teachers were asked to report the regularity of implementation during the academic year. Some implemented the programme weekly, whilst others once or twice per month. This might have a different effect on the programme impact. Also, the two variables (sex and age) referring to students' characteristics were not considered. As a result, regressions considered these variables. Given the gender, the age of the student and their participation in Philosophy for Children sessions, would it be possible to predict their post-test results in critical thinking?

Two models were created. The first one was based on a regression with only one step, whilst the second one with two steps. Table 11.8 presents the variables included in each of the two models. The results of the two models showed that no variable included in the model could explain sufficiently any change in the post tests results.

Table 11.8. Variables and variance explained for the two models for Critical Thinking Skills

Models	Predictors	R

		Square
1	CT Pre-test Performance (School Level), Age, Sex	0.103
2	CT Pre-test Performance (School Level), Age, Sex, Frequency of Sessions the last academic year, Number of Years, Intention to Treat	0.134

Table 11.9. Regression for Critical Thinking Skills (Beta Standardised Coefficients)

	Model 1	Model 2
Sex	0.096	0.093
Age	-0.004	-0.003
CT Pre-test Performance (School Level)	-0.041	0.001
Number of Years		0.141
Frequency of Sessions the last academic year		-0.136
Intention to Treat		0.008

The models (Table 11.19) showed that girls tended to have slightly higher post-test results than boys. This could be sample dependent findings since the sample was not randomly selected. However, it might be the case that the girls performed slightly better than boys for other reasons. Instead of supporting that the girls might have more critical thinking than boys, it might be the case that the assessment introduced some type of bias. While the assessment was constructed, the topics in the thinking problems and the characters were carefully selected in order not to be of an interest of a specific sex only. Equal numbers of male and female characters appear in the assessment problems. Considering the fact that there is evidence supporting that girls usually perform better than boys in reading (Marks, 2008) and the thinking problems were indeed linguistic in this assessment, girls may perform better. However, the assessments were multiple-choice questions and closed items which – according to evidence - are usually in favour of boys (Beller & Gafni, 2000; Bolger & Kellaghan, 1990; Yip, Chiu & Ho, 2004).

Moreover, younger Year 5 students tended to perform better than their older classmates. The model presents a slightly negative relationship between the performance of the students and their age, since the performance of the students in both groups (intervention and control) was slightly lower in the post-test. It is likely that as the students grow older their critical thinking seems to deteriorate. This might be due to various factors, such as longer involvement in higher education and formal

education. It might be the case that the students conform to ways of thinking and given answers.

Lipman (2003) also noticed that even though the students are naturally curious, they do not demonstrate critical thinking when they are in the university. He introduced P4C curriculum as a way to help the students develop this natural curiosity. Consequently, even though this research does not provide evidence that P4C develops the critical thinking of students, it provides some indicators that critical thinking is not fostered by schooling and it might be restricted when the students become older. This deterioration may appear due to other factors. For example, items in the post-test may have been more difficult than those in the pre-test. To summarise, the contribution of both the factors 'sex' and 'age' is small in the models. There were also three variables related to the P4C implementation.

Since the mean score of the pre-test for the whole schools was used as a baseline assessment, it is expectable that it could not accurately predict the individual post-test performance. There is only a weak relationship between the two. This relationship is negative because the intervention group, which performed better than the comparison group in the pre-test, performed worse than the comparison group in the post-test.

The three variables related to P4C which were included in the model (intention to treat, frequency of the sessions and number of the years that the school was involved in the programme) do not predict the performance of the students in the critical thinking problems. Additionally, these three variables provide contradictory results. Even though the students who attend a school involved in P4C perform better than the students of the comparison group (and in fact the students whose school have been involved for more years perform even better), the frequency of the P4C sessions seem to be slightly negatively correlated to the performance of the students. However, it has to be noted that the school which was included in the intervention group and had the highest number in the years of implementation, it stopped implemented the P4C during that year. Hence, it was included in the intervention group with intention to the treat analysis but the frequency of P4C sessions was zero.

There is some collinearity between the frequency of the sessions and the intention to treat. However, these variables are not identical. There was a school which stopped implemented P4C during that academic year and even though it is the school with the most years of implementation, it is included in the intervention group

(intention to treat) with the frequency of sessions being zero. Similarly, one of the classes in one of the comparison group started doing P4C sessions. Despite the negative relationship between the frequency of implementation and critical thinking, I do not argue that irregular implementation of the programme would be more effective than frequent implementation. Currently, the programme has a negative impact on critical thinking skills. This is the reason why the regression suggested a negative relationship between frequency of implementation and critical thinking skills score. However, if the programme is adjusted and included explicit teaching of the skills, a regular implementation of the programme is recommended.

11.4. Creativity: Results

The following section discusses the P4C impact on creativity. Before this discussion, however, I judged necessary to include some challenges of the marking of the creativity assessment. In the methods chapter, I presented the process followed for marking the creativity activities. However, I chose to include these challenges of marking in the results section, because in practice the process had specific challenged. The implementation of the methods was not a straightforward process and the way these assessments were marked might have potentially influenced or slightly biased the results presented in this section.

The first activity of 'uses of objects' had two main challenges when marked. The handwriting of the children was sometimes too difficult to be read. This process was impeded particularly when the spelling might be wrong. However, four more assessors were engaged in this process and native speakers helped me, so we tried to read as many responses as possible.

Only if five people (including myself) could not read the responses, an answer was accepted as illegible. However, there were a few cases that more than most of the raters could not read an answer and one of the raters managed to identify what was written. Table 11.10 presents the percentage of answers which were marked as illegible in comparison to the overall number of answers which was judged invalid. The reasons that answers were judged invalid were presented in chapter 8.

Table 11.10. Frequency of Answers for the first creativity activity.

Uses of Objects	Pre-test	Post-test
N of Illegible Responses	23	21
N of Invalid Answers	725	418
Percentage of Invalid Answers which were illegible	3.17%	5.02%

Total N of Valid Answers	4,799	4,598
Total N of answers	5,524	5.016
Percentage of Invalid Numbers	13 %	9%

It has to be noted that only a small percentage of the responses judged as invalid were illegible responses. Therefore, in most of the cases one of the five raters managed to read the responses. However, it becomes obvious that the pre-test had almost double responses been judged invalid (N = 725) compared to the post-test (N = 418). Possible interpretations could explain this. First, the pre-test and the post-test used different objects. Thus, in pencils it is likely that the students tended to mention more school objects which were not related to the use of a pencil. As a result, on the one hand the one assessment might have had more invalid responses because of the topic. On the other hand, there might be some unconscious bias from myself in the post-test and I might have graded more lenient without actually realising this.

In Chapter 8, I explained the criteria used to judge a response as invalid. However, I also examined whether there was unconscious bias. At this point, it is important to examine the relationship between the fluency score and the invalid responses provided by the students. The correlation for the pre-test was r=0.098 and r =0.040 for the post-test. Therefore, there is no bias which disadvantages a particular type of students (high or low performers) in the process of judging answers as invalid.

Table 11.11. Frequency of Questionnaires in relation to the type of responses in creativity activity 1

Type of responses in the Questionnaire	Pre-Test	Post-Test
Only Valid Responses	476	527
1 Invalid Response	194	126
2 Invalid Responses	65	38
More than 2 invalid responses	82	47

The marking of the second activity had also the challenge of reading the handwriting of the pupils. Table 11.12 shows that less than 1% of the participating students gave illegible answers for the second activity.

Table 11.12. Frequency of Illegible Answers for the second creativity activity

Abstractness of Title	Pre-test	Post-test
Illegible Answers	7	8

11.4.1. Missing Data

This section reports the missing data of this assessment (Table 11.13). The missing data of the creativity activities was less than 2.5 %. The pupils generally responded to these activities, despite the fact that it required them to respond to open-ended questions and there were no assessment consequences. This could be explained by the fact that these questions appeared at the beginning of the assessment and therefore the students were not tired.

Table 11.13. Frequency of Missing Data in the creativity activities

	Pre-test			Post-test		
	Activity 1	Abstractness	Resistance	Activity 1	Abstractness	Resistance
Valid	804	798	807	733	720	719
Missing Data	13	19	10	5	18	19
Total	817	817	817	738	738	738

The low percentage of missing data is very interesting. I expected a higher percentage of missing data because the assessment required the generation of responses with no clearly stated purpose or reward. Amabile (1985) discussed the relationship between motivation and creativity. However, in this activity the students did not seem to have a clear extrinsic or intrinsic motivation to complete these activities. However, students generated responses. If creativity is a purposeful thinking which aims to respond to an existing situation or problem, the reasons why the students completed the assessment may be questioned. These assessments were not linked to an authentic situation or a problem. The low percentage of missing data might suggest the impact of the environment of creativity. Students were asked by their teachers to complete these tasks and therefore they proceeded to the generation of responses.

11.5. Creativity Skills

11.5.1. Relationship between Sub-categories

It was important to determine the overall score of the creativity in order to respond to the research question. Before calculating the overall score of the creativity, the relationships between the fluency, flexibility and two prevalence indicators were calculated because there is an assumption that fluency and originality scores in the Torrance Test of Creative Thinking are related (Kim, 2006). Torrance Tests for

Creative Thinking did not include flexibility (Torrance, Ball & Safter, 2008). Torrance removed the flexibility scale from his divergent thinking assessment because it was too highly correlated to fluency scores (Plucker & Makel, 2010).

For this thesis, if two sub-scales were highly correlated, they should have not been included in the calculation of the overall score. The overall score of creativity was calculated as a sum of different components. Thus, skills which were too highly correlated to each other, they both added the same information. If both of them were included in the overall sum, that element would have had double weighting in the final calculation.

The results of the correlation of this thesis confirmed that there is a considerable overlap between flexibility and fluency (r = 0.85 in the pre-test and r=0.70 in the post-test). Therefore, there was no need to include both fluency and flexibility because they were both measuring actually the same thing. This might be due to the fact that as the chapter of grading suggested, fluency was graded 'almost' without any qualitative evaluation. However, it involved qualitative evaluation to some extent and therefore this could be the reason why fluency and flexibility were found to be related.

Similarly, the sum prevalence, which was the first indicator to measure prevalence, was also found highly correlated with fluency. Similarly, prevalence sum was also highly correlated with flexibility. As a result not all the subscales were necessary for the sum for creativity, because some of them did not offer additional information by measuring the same exact element of the construct.

Table 11.14. Matrix with inter-item correlations for the pre-test

	Fluency	Flexibility	Prevalence Sum	Maximum Value	Abstractness to Title	Resistance to Premature Closure
Fluency	1	0.850	0.835	0.485	0.122	0.142
Flexibility			0.994	0.639	0.165	0.137
Prevalence Sum				0.635	0.165	0.135
Maximum Value					0.187	0.105
Abstractness						0.215

to Title						
Resistance to Premature Closure						1

Table 11.15. Matrix with inter-item correlations for the post-test

	Fluency	Flexibility	Prevalence Sum	Maximum Value	Abstractness to Title	Resistance to Premature Closure
Fluency	1	0.701	0.687	0.432	0.004	0.047
Flexibility			0.993	0.651	-0.006	0.002
Prevalence Sum				0.633	-0.006	-0.008
Maximum Value					0.029	0.035
Abstractness to Title						0.072
Resistance to Premature Closure						1

Therefore, since fluency, flexibility and prevalence sum were found to be highly correlated, it was decided that only one of these would be included in the calculation of creativity. Since there was an alternative indicator for prevalence and originality, which was the maximum value, the prevalence was excluded. However, the choice between fluency and flexibility might seem a bit arbitrary. Nevertheless, there was an existing recommendation from Getzels and Jackson (1962). They suggested that the 'Uses for Things' assessment is scored for the number of different uses suggested for that object and the number of uncommon uses. Therefore, they suggested only these two scores. The first one clearly matches to the definition of flexibility. Thus, flexibility and maximum value were used for the calculation of the overall score of the creativity. These two variables were also combined with the two variables from the second activity.

Torrance removed flexibility from the scoring of creativity, whilst Getzel and Jackson suggested including the flexibility and excluding the fluency. Based on the

definition of each of the two aspects of creativity, I considered flexibility as a variable which is more valuable than simply fluency, which referred only to the quantity of creativity with no in-depth quality evaluation.

It should be noted that the variable 'maximum value' is that not highly correlated with the fluency score. Whilst the variable 'prevalence sum' takes into consideration both the prevalence and the number of the answers provided, the 'maximum value' variable only considers the most rare and innovative answer provided by an individual without considering the number of responses provided. The data did not suggest a relationship between the variables of fluency and 'maximum value', which clearly indicates that the number of responses is not highly correlated with the innovation of these ideas. The implication that this finding might have on a real-life context is that the individuals who produce the most do not necessarily produce the most innovative products.

11.5.2. Calculation of Creativity Overall Score

As it becomes apparent from the previous section, the calculation of creativity was not pre-decided. The fact that the two sub-sections of creativity (fluency and flexibility) were highly correlated was expected based on the literature. However, it was examined also based on the data of the specific study. Since flexibility and fluency were particularly highly correlated, this revealed that both variables provide the same information. Therefore, only one was needed for the calculation of creativity. Similarly, the 'maximum value' variable was also more informative than the 'prevalence sum' and hence the first was included as an indication of innovation instead of the latter.

As a result, creativity overall was calculated using the following formula

$$Creativity_{Overall} = \frac{Flexibility + Maximum\,Value_{\,Sample\,Adj} + Resistance\ to\ Pr.Closure + Abstractness}{4}$$

However, since each of these variables was measured on a different scale, they were first turned into z-scores in order to enable the calculation of the Creativity variable.

11.5.3. The impact of Philosophy for Children on creativity

The impact on creativity was judged based on the intention to the treat analysis, as in the case of critical thinking analysis. Therefore, the one class which started

implementing P4C despite being in the comparison group and the one school which stopped implemented the programme were categorised in their initial groups.

Table 11.16. Impact of Philosophy for Children on Creativity

	Pre-test			Post-test		
Overall Creativity	Sample (N)	Mean	Standard Deviation	Sample (N)	Mean	Standard Deviation
P4C Group	547	0.04	0.64	495	-0.01	0.61
Comparison group	270	-0.08	0.66	243	0.02	0.54
Effect size	0.19			-0.05		
Effect size	**-0.24**					
NNTD	58					

The impact that P4C has on creativity appears to be slightly negative. In order to disturb this finding, 56 counterfactual cases are needed. The participants who dropped out before the post-test were more than the number of counterfactual cases. This suggests that the results are provisional and they could have been different if there was no dropout.

Even if P4C does not develop creative thinking it might reduce the dark side of creativity. There was no extensive evaluation of this. However, it might be interesting to see how the creativity is affected by P4C, since Lipman (2003) also argued that P4C develops the caring thinking.

At this point, the two hypotheses set in the literature chapter are tested. The first one referred to the recent use of objects. Guilford (1967, p.327) argued that recent use of objects in their common and conventional uses made more difficult to think of unconventional uses of these objects. During the pre-test, the students were requested to suggest uses for pencils which are more commonly used object in the pupils' lives compared to bricks. This finding is not confirmed by the data of this thesis, since the comparison group did perform worse when suggesting a use of a pencil than a brick. However, this might explain the higher number of invalid answers provided for pencils compared to a brick. The familiarity of the object might have led to inclusion of irrelevant responses.

The second hypothesis was related to the age of the participants. Torrance (1962) reported that fourth graders produced less compared to the other grades. He also

explained that some of these students will lose their creative growth rather permanently and he discussed different explanations for this decrease in creative development, such as physiological explanations. The age factor did not seem to play a role in the performance of the two groups, since the one group improved their performance and the other decreased their performance as they became older.

11.6. The impact of Philosophy for Children on different aspects of creativity

As in the case of critical thinking, the P4C impact on each of the skills included in the creativity construct was reported separately (Tables 11.17 -11.22). This is due to the fact that some of these aspects were not included in the calculation of the creativity overall because as it was previously mentioned, they were highly correlated to each other.

Table 11.17. Impact of Philosophy for Children on Fluency

	Pre-test			Post-test		
Fluency	N	Mean	Standard Deviation	N	Mean	Standard Deviation
P4C Group	547	6.15	4.19	495	6.06	4.73
Comparison group	270	5.34	3.55	243	6.60	4.23
Effect size	0.20			-0.12		
Effect size (pre-test and post test)	**-0.33**					
NNTD	87					

Table 11.18. Impact of Philosophy for Children on Flexibility

	Pre-test			Post-test		
Flexibility	N	Mean	Standard Deviation	N	Mean	Standard Deviation
P4C Group	547	3.60	2.61	495	2.91	2.28
Comparison group	270	2.94	2.26	243	3.15	2.04
Effect size	0.26			-0.10		
Effect size (pre-test and post-test)	**-0.17**					
NNTD	41					

Concerning fluency (Table 11.17), the negative effect size is mainly due to the increase of the score comparison group. The average score in the intervention group was about

6 in both the pre-test and the post-test. This means that on average students in the intervention group reported six uses for each object in both tests.

Table 11.18 presents the impact of the programme on flexibility. The programme has a negative impact on flexibility. After the participation in P4C sessions, students reduced their score in flexibility. In fact, this might be a result of the intervention. As Dewey (1933) suggested reflective thinking is not merely a sequence of ideas but a 'con-sequence'. Since P4C aims to increase reflective thinking, this might decrease flexibility. I argue that flexibility is a type of thinking moving towards different directions and this is contradictory with a purposeful reflective thinking. Therefore, if the latter is increased via participation in P4C sessions, then the first might deteriorate.

Table 11.19. Impact of Philosophy for Children on Prevalence

	Pre-test			Post-test		
Sum Prevalence (adjusted for difference in sample size)	N	Mean	Standard Deviation	N	Mean	Standard Deviation
P4C Group	547	2.48	2.29	495	1.82	2.02
Comparison group	270	1.87	1.99	243	1.99	1.82
Effect size	0.28			-0.08		
Effect size (pre-test and post-test)	**-0.38**					
NNTD	92					

Table 11.20. Impact of Philosophy for Children on Innovation (Maximum Value)

	Pre-test			Post-test		
Maximum Value (adjusted for difference in sample size)	N	Mean	Standard Deviation	N	Mean	Standard Deviation
P4C Group	547	0.73	0.38	495	0.66	0.41
Comparison group	270	0.68	0.38	243	0.75	0.36
Effect size	0.13			-0.23		
Effect size (pre-test and post-test)	**-0.36**					
NNTD	87					

P4C has a negative impact on prevalence and maximum value (Tables 11.19 and 11.20). The intervention group performed better than the comparison group in the pre-test and worse in the post-test. The intervention group appears to provide less innovative responses in the post-test compared to the pre-test. However, these two

variables are sample-dependent. This might be a result of a dialogue. Students in a community of enquiry exchange ideas and therefore it is likely that the ideas they mentioned in the post-test were similar to their classmates' ideas.

This can be interpreted in a positive way. Instead of perceiving that community of enquiry decreased innovation, it is likely that this finding suggests two benefits. First, students start sharing some ideas and this created homogeneity in the ideas they mentioned. Secondly, they may have collaboratively co-created some innovative ideas. Even though these do not appear innovative in a normative sample-dependent assessment of innovation, their ideas may be innovative if compared to external criteria or groups.

It should be clarified that the negative effect sizes on the skills of fluency, prevalence sum and maximum value (Tables 11.17 -11.20) are considered generally trustworthy, because the NNTD was bigger than the number of participants who dropped out of the study (79 cases dropped out). Therefore, these findings are trustworthy.

The findings of this thesis can be directly compared to these reported by Pourtaghi, Hosseini & Hejazi (2014) who used the Torrance Test in order to identify the P4C impact on creativity. Their findings suggested that P4C has a big positive impact on all four domains of creativity included in the Torrance Test (fluency, flexibility, innovation and elaboration). Their sample was not big, and this could explain the big effect sizes. However, the effect sizes were also confirmed by the calculation of effect sizes in the previous chapter of this thesis. This thesis used only the performance of the two groups in the post-test to calculate the effect sizes. Therefore, it has to be mentioned that big effect sizes were found despite the fact that there is pre-test imbalance with the comparison group being ahead of the intervention group.

Table 11.21. Impact of Philosophy for Children on Resistance to Premature Closure

	Pre-test			Post-test		
Resistance to Premature Closure	N	Mean	Standard Deviation	N	Mean	Standard Deviation
P4C Group	547	1.77	0.92	495	2.22	0.88
Comparison group	270	1.66	1.01	243	2.12	0.86
Effect size	0.12			0.11		
Effect size (considering both pre-test and post-test	**-0.01**					

performance)	
NNTD	2

Table 11.21. presents the impact of the programme on resistance to premature closure. Both groups developed this skill and performed better in the assessment at the end of the academic year. The programme had no impact on this skills. The improvement of scores for both groups suggests that there is probably a different factor which facilitates the development of this creativity skill for the students. This might be an aspect of schooling. Students might have also practiced to the test and performed better in the second assessment. It has to be noted that the activity in the assessment at the end of the school year was similar to the one in the assessment at the beginning of the academic year and therefore practice to the test was possible.

Table 11.22. Impact of Philosophy for Children on Abstractness of Title

	Pre-test			Post-test		
Abstractness of title	N	Mean	Standard Deviation	N	Mean	Standard Deviation
P4C Group	547	0.82	0.77	495	0.88	0.79
Comparison group	270	0.81	0.77	243	0.81	0.66
Effect size	0.01			0.09		
Effect size (considering both pre and post-test performance)	**0.08**					
NNTD	19					

P4C has a slightly positive impact on ‘the abstractness of the title’ (Table 11.22). This area was not examined by Pourtaghi, Hosseini & Hejazi (2014) and therefore the results cannot be compared. P4C intervention involves discussion on philosophical topics which use abstract vocabulary and this might explain this finding. The marking of the title appears to measure more vocabulary enrichment with abstract concepts. While simple titles are marked with 0, more complicated titles with adjectives are marked with 1 which might be mainly vocabulary enrichment rather than a creativity element. Abstract concepts are marked with 2. As mentioned in Chapter 2, P4C emphasises defining concepts (Bassiri & Vaidya, 2013), which is one main element of philosophising. This might explain why the students in the P4C group appear to score higher in this sub-section of creativity. It is likely that they were more familiar with abstract concepts and they might have used them more often. In that sense though, P4C

intervention is effective. Familiarising pupils with these concepts and enhancing their understanding are definitely included in the targets of the intervention.

Even though this finding might seem related to the linguistic skills of students, this is not necessary. Dewey (1933, p.241) discussed how enlarging students' vocabulary facilitates their thinking. According to him, 'paucity of vocabulary' is one of the aspects that 'tend to shut down the area of mental vision'. Therefore, by clarifying concepts can link to ideas and thinking.

11.7. Regression for Creativity Performance

As mentioned in an earlier chapter, the fact that there were schools which implemented P4C for a different number of years was used as an opportunity to create models for the critical thinking and creativity performance.

The same variables which were included in the models for critical thinking were also included in the models for creativity. Similarly, neither of the two models was able to predict the creativity performance of the students. When the pre-test performance was uses as a predictor for the post-test performance in critical thinking problems, it was not found related to it. This can be interpreted as lack of sensitivity in the variable, since the pre-test performance on a school-level was used instead of the performance on an individual level.

Table 11.23. Variables and variance explained for the two models of Creativity

Models	Predictors	R Square
1	Creativity Pre-test Overall Performance (School Level), Age, Sex	0.05
2	Creativity Pre-test Performance (School Level), Age, Sex, Frequency of Sessions the last academic year, Number of Years, Intention to Treat	0.06

Table 11.24. Regressions for Creativity (Beta Standardised Coefficients)

	Model 1	Model 2
Sex	0.048	0.046
Age	0.014	0.020
Creativity Pre-test Performance (School Level)	0.219	0.252
Number of Years		0.036
Frequency of Sessions the last academic year		-0.117
Intention to Treat		-0.013

In the case of creativity, the pre-test performance (school mean) was positively correlated to the performance of students in the post-test (Table 11.24). This might be due to the fact that the creativity activities used in the pre-test and post-test were similar and students practiced to the test. This was not the case in critical thinking problems. Those questions evaluated the same critical thinking skill, but the content of the problems differed between pre-test and post-test.

Females performed slightly better in the creativity assessment, as in the case of critical thinking. There was no relationship with age. Based on what Torrance (1962) reported the students might gradually overcome the fourth-grade slump and they may have started improving their creativity again. The implementation of P4C appeared to have a slightly negative impact (as the effect size showed) and creativity was reduced marginally. A negative relation was found between the frequency of P4C sessions and the creativity. Consequently, females performed slightly better than males both in critical thinking and creativity assessments. It has to be clarified though that the tasks were linguistic and required reading and writing. Girls tend to perform better in this type of tasks in general.

11.8. Summarising and Interpreting the Results

This chapter presented the findings of the comparative evaluation study conducted. That trial examined the impact of P4C on critical thinking and creativity. There was no study conducted which evaluated all these different skills of critical thinking. Similarly, there was no large-scale evaluation study which assessed the impact on creativity. Therefore, the findings of this thesis were very important to shed light on the programme effectiveness on thinking skills.

P4C did not have an impact on critical thinking. Overall, P4C was found to have no impact on critical thinking skills. Nevertheless, the findings of this thesis are tentative. This is due to the fact that the NNTD was bigger than the number of participants who dropped out. Therefore, without attrition the findings of the study might have been different and no strong claims can be made.

The findings of this study demonstrate that when there is implicit teaching of critical thinking, these skills do not improve. It is likely that the students either had no opportunity to learn and practice these skills or this learning took place in a very implicit way. Lipman's novels attempted to teach these in a more direct way. Lipman's

structured approach might have led to better results. There is no evidence to suggest this. However, it has been argued that explicit teaching of critical thinking skills is a promising approach of developing these skills in higher education (El-Soufi & See, 2019). Furthermore, based on the meta-analysis of findings for the development of critical thinking at different educational levels (Abrami et al., 2015), the direct teaching of critical thinking skills was found to be more effective than teaching them with the immersion approach. Therefore, if critical thinking skills are expected to change, there should be direct teaching and practice of these skills. Currently, in P4C sessions this does not seem to be the case and this might explain why the only area on which P4C had impact was related to vocabulary.

P4C was found to have negative impact on creativity. Even though fluency, flexibility and the sum of prevalence were highly correlated and measured the same skill, the programme was also found to have negative impact on the innovation (as prevalence and maximum value). For the skills of fluency and innovation, the NNTD was smaller than the number of drop outs. Therefore, it can be confidently said that there is negative impact of the programme on the performance of students in the creativity activities. It is likely that P4C develops the purposeful thinking which is not associated to the divergent thinking. This can explain the negative impact of the programme on divergent thinking skills.

There was a small positive impact of the programme on the abstractness of title the pupils used on the creativity activity. It is likely that this aspect of the assessments was more closely related to the elements of the programme and this is why it had a positive impact on it. Students discuss ‘big ideas’ during the P4C sessions, many of which are abstract concepts. However, this aspect of creativity did not provide as trustworthy findings as the other elements of creativity, since the NNTD was bigger than the number of the participants drop out. P4C might have caused vocabulary acquisition. Hence, students’ performance increased in the use of abstract concepts because students learnt and used them during the P4C sessions.

12. Results of the Secondary Data Analysis: The Philosophy for Children Impact on Attainment

As became apparent in the systematic literature review, there were a few studies which examined the P4C impact on attainment. However, these studies did not give a consistent overview about the effectiveness of the programme on attainment. Six positive effect sizes were reported in the studies regarding students' literacy skills. However, there were also three negative effect sizes. Concerning Maths, the evidence was even less to support any conclusions. Hence, there was contradicting evidence and no evidence about the impact of the programme long-term.

For this reason, the fourth research question of this thesis concerned the impact of P4C on attainment, particularly when implemented for four successive years in Key Stage 2. Key Stage 1 results were used as a baseline assessment and Key Stage 2 as a post-test assessment. The areas examined were reading, writing and mathematics. Key Stage 2 results from the year 2015 were used and therefore P4C was implemented from 2011-2015 at least in the schools in the intervention group (P4C schools).

The impact of P4C on attainment is crucial. Despite Lipman's emphasis on thinking skills, I argue that he would agree with the presentation of the impact of P4C on attainment. Specifically, in the publication of Bierman report based on the first P4C project, Lipman (1976; 1982) claimed significant impact on reading skills for the P4C group. Despite the fact that attainment was reported then, currently P4C adherents focus more on the development of thinking. For example, SAPERE training emphasises the development of 4 C's (critical, creative, collaborative and caring thinking) in their training (SAPERE, 2015c).

Indeed, P4C is a noteworthy intervention because it claims to improve thinking. Thinking should be the main focus of P4C. The improvement of attainment can and should be included in the presentation of the impact of the programme overall without being its main focus, since the impact on reading was an examined area from the beginning of P4C and an important aspect of schooling.

12.1. Results: Impact on Attainment

According to the Table 12.1., the P4C group performed better than the comparison group in all three subjects in Key Stage 1 assessments. This might raise questions about the type of schools which sign up for P4C training. Even though there might be no causation between school characteristics and the decision to sign up for P4C, it is

likely that there is some correlation. Based on the table 12.1., the input of schools which received P4C training was students with higher attainment. Similarly, based on the table 5.4 presented in an earlier chapter, P4C schools in the comparative evaluation study had low proportion of SEN students.

Students in the intervention group also performed better in Key Stage 2 results in 2015. However, the performance gap between the two groups closed. Therefore, the effect sizes with both pre-test and post-test comparison appear slightly negative. If there was only consideration of the post-test performance, the effect sizes would have been positive because the P4C schools still performed slightly better than the control schools in Key Stage 2 assessments of reading, Maths and writing. However, this image would have been distorting since they were already performing better in the baseline assessments (Key Stage 1).

Table 12.1. Impact of Philosophy for Children on Attainment

		Key Stage 1			Key Stage 2		
Years 2011-2015		N pupils	Mean	Standard Deviation	N pupils	Mean	Standard Deviation
Reading	P4C Schools	2,735	0.12	0.97	2,735	0.07	0.99
	Comparison Schools	560,499	0.00	1.00	560,499	0.00	1.00
	Effect Size	0.12			0.07		
	Difference	**-0.05**					
Writing/GPS Fine	P4C Schools	2,735	0.13	0.99	2,735	0.09	0.99
	Comparison Schools	560,499	0.00	1.00	560,499	0.00	1.00
	Effect Size	0.13			0.09		
	Difference	**-0.04**					
Maths	P4C Schools	2,735	0.10	0.99	2,735	0.06	0.99
	Comparison Schools	560,499	0.00	1.00	560,499	0.00	1.00
	Effect Size	0.10			0.06		
	Difference	**-0.04**					

There is a likely interpretation for this finding. The main aim of P4C is not the improvement of attainment. It is likely that the schools in the comparison group allocate more time on interventions which specifically aimed to the improvement of students' attainment.

12.2. Results: Impact on Disadvantaged Students' Attainment

Another question set was whether P4C has more positive impact on the attainment of disadvantaged students. Students eligible for FSM the last 6 years improved their performance in reading and writing when they receive P4C sessions compared to those who did not (Table 12.2.). A slightly negative effect size was found for Maths.

Table 12.2. Impact of Philosophy for Children on Attainment of Students Eligible for Free School Meals (Ever FSM6)

Years 2011-2015		N pupils (pre-test)	Pre-test Key Stage 1 (Mean)	Standard Deviation	N pupils (post-test)	Post-test Key Stage 2 (Mean)	Standard Deviation
Reading	P4C Schools	799	-0.24	1.02	799	-0.15	1.04
	Control Schools	173,819	-0.37	1.04	173,819	-0.31	1.07
	Effect Size	0.13			0.15		
	0.03						
Writing/ GPS Fine	P4C Schools	799	-0.26	1.02	799	-0.13	1.00
	Control Schools	173,819	-0.37	1.02	173,819	-0.30	1.03
	Effect Size	0.11			0.16		
	0.06						
Maths	P4C Schools	799	-0.24	1.04	799	-0.23	0.98
	Control Schools	173,819	-0.36	1.02	173,819	-0.32	0.97
	Effect Size	0.12			0.09		
	-0.03						

There is a likely interpretation for this result. Students might improve their performance in reading and writing due to the characteristics of the programme. P4C involves students in dialogue and therefore the linguistics skills and their ability to write arguments can be increased. On the contrary, there is no direct or indirect association with Maths-related skills. This might explain why the programme did not appear to have a positive impact on maths.

FSM students might increase their linguistic skills because of the participation in P4C, whilst the group overall did not (see Table 12.1). This is because P4C sessions

give the opportunity to every student to express their opinion and build arguments since there is no right or wrong answer in this dialogue. The finding of this thesis confirms the argument that 'language is both socially developed and socially situated' (Pring, 1980, p.14). In other subjects, it is likely that students with lower attainment participate less in order to avoid giving a wrong response. During P4C sessions each opinion is respected and each student can contribute to the dialogue with their own experiences. No pre-requisite knowledge is necessary for the participation in the dialogue. Students from low socio-economic background can contribute to this type of discussion and develop their linguistic skills.

12.3. Discussion

The analysis of secondary data of this thesis showed that P4C had a slightly negative impact on students' attainment, but a slightly positive impact on the literacy skills of FSM students. The findings of this thesis are only partially in agreement with the findings of previously published studies. The effect size calculated for Youssef, Campbell and Tangen (2016) showed a small positive impact of P4C on students' reading comprehension. Gorard, Siddiqui and See (2017) reported small effect sizes (around +0.1) for reading and Maths for all the students, and larger impact on progress scores for reading (+0.29), writing (+0.17) and Maths (+0.20) for disadvantaged students who were eligible for Free School Meals. The researchers used Key Stage 1 points as a baseline assessment and Key Stage 2 fine scores as a post-test. That study was a randomised controlled trial with 48 schools in England. The randomised controlled trial considered the P4C impact on reading, writing and Maths after an intervention which lasted for approximately one academic year.

The results of this thesis are in agreement with the findings of Gorard, Siddiqui and See (2017) when reading and writing of disadvantaged students are concerned. Both studies found a positive impact for FSM students. However, this thesis does not report effect sizes which are as big as the conducted randomised controlled trial. The benefit of the programme on students eligible for FSM reported by this thesis is in line with other studies which found the disadvantaged students to be more benefited from the programme than the other students in the intervention group and the comparison group (Colom et al., 2014; Gorard, Siddiqui & See, 2015; Sasseville, 1994; Stokell, Swift & Anderson, 2017). The findings of this thesis are in line with Jo (2001), who reported that students who participate in P4C sessions performed better at meaning

construction compared to the comparison group. This means that there might be an improvement of language skills during P4C sessions.

An unpublished report by Swain, Cara and Litster (2014) also evaluated P4C impact on reading. In that study, the intervention group consisted of 236 students, whilst the control group of 250 students and their intervention was short-term (approximately 9 weeks and 9 hours of philosophy). The researchers reported no positive impact of the intervention on the reading skills of the students. The reading scores of both groups increased in the post-test. In that study, there was no examination of the impact on FSM students. Maybe the researchers did not report this separately, because there were only 75 FSM students in the intervention group and 87 in the control group. However, their findings can be considered consistent to the findings of this thesis which found no positive impact on the reading of the students in the intervention group. Despite their intervention being short-term, their findings are consistent with these of Gorard, Siddiqui and See (2017) and the results of this thesis, which adopted a long-term approach.

To summarise, based on the evidence overall, there are mixed findings about the impact of the programme on attainment. Earlier in this thesis (Table 10.3.), it became apparent that there is more positive than negative evidence related to the impact of the programme on literacy. Furthermore, the evidence overall suggests that the attainment of disadvantaged students is developed when P4C is implemented. The analysis of secondary data presented in this thesis showed that the positive impact of the programme on FSM students' literacy skills can be observed after 4-years P4C implementation. Therefore, long-term implementation of the programme is recommended to close the attainment gap between advantaged and disadvantaged students.

Based on the secondary data analysis of this thesis and the only large-scale randomised trial which examined the impact of the programme on attainment, it is supported that P4C closes the attainment gap. However, it should be clarified that these two sources of evidence suggest two distinct ways of closing the attainment gap. The analysis of secondary data of this thesis suggested that the attainment gap closes because the disadvantaged students slightly improve their literacy whilst the advantaged students in the classrooms decrease their performance. Thus, the mean performance of the group overall decreased. Evidence from Gorard, Siddiqui and See

(2017) suggested that both advantaged and disadvantaged students increased their literacy. However, there were bigger gains in the literacy of disadvantaged students.

Consequently, both studies suggested that the attainment gap between advantaged and disadvantaged students closes when P4C is implemented. It can be questioned whether it is acceptable and ethical to close the gap by impeding the progress and increase in scores for the advantaged students.

13. Limitations

This chapter discusses the main limitations of this study. It discusses each of the three methods (the systematic literature review, the comparative evaluation study and the secondary data analysis of National Pupil Database) separately from each other and in relation to the research question each of those aimed to answer.

13.1. Systematic Literature Review

This thesis used effect sizes to examine the effectiveness of P4C. Simpson (2017) suggested that the effect sizes can be misleading, and he appeared concerned about their use. Specifically, he argued that the comparison of effect sizes between different studies can be ambiguous for various reasons. Firstly, there are differences between the approaches implemented for the comparison groups in each of the studies. Some of the studies provide no treatment to the comparison group, while others provide alternative or placebo treatments. The systematic literature review of this study did not examine what the comparison group received. Comparison groups across studies were assumed to receive regular teaching. Nevertheless, this might not have been the case. Furthermore, what is considered regular practice is not homogeneous across schools.

The calculation of the effect sizes uses descriptive statistics reported for the comparison group and therefore the quality of what the comparison group received plays a role in the final findings. In the case of P4C, a comparison group which receives an authoritative teaching would be expected to have different post-test results compared to a group which receives a different dialogic intervention.

Simpson (2017) also claimed that different measurement tools and populations chosen by researchers affect the effect sizes of the studies. A focused measurement tool on the trait of the intervention can make the intervention appear more effective than what it is. In the case of this systematic review, Säre, Luik, and Tulviste (2016) study is an obvious example of this case with a measurement tool focused on P4C elements. However, there was no detailed presentation of the measurement tools in the systematic literature review of this thesis.

There is the threat of diffusion between the two groups (Gorard, 2001, p.139). This phenomenon could be more intense when participants from both groups attend the same school. Then, it is likely that the intervention and information about it are shared between the control and intervention group. In simple words, the comparison group

might not be clean of the intervention and this can lead to a smaller effect size. The systematic literature review could not control or examine potential diffusion.

Although comparing effect sizes from different studies might lead to misleading results, the aim of this systematic literature review was not the aggregation of the effect sizes, as a meta-analysis would do. The aim of the systematic review was the overall programme evaluation by considering all the existing evidence. Thus, even if an effect size does not represent the real magnitude of the effect, the consideration of all the effect sizes suggested the effectiveness of the programme. The magnitude of the reported effect sizes on specific skills might have been distorted by factors, such as the measurement tools or comparison group interventions. It has already been discussed in Chapter 10 that bigger effect sizes were found for studies graded with less stars.

13.2. Comparative Evaluation Study

Concerning the comparative evaluation study, the sample is not random and there is no random allocation of participants or schools within the groups. Even though the latter characteristic did not lead to an initial imbalance in the performance of the two groups, it would be difficult to generalise the results to the population of English schools.

Even though these results clearly demonstrate that the intervention group did not perform better than the comparison group, they do not necessarily mean that the intervention is ineffective. There might be different factors that distort the impact that P4C might have on critical thinking. The research project lasted only one academic year. The students did not have the opportunity to be involved in many P4C sessions during this time. Some of the schools implemented P4C twice per month and there was no school which implemented the intervention more than 2 times per week for 30 minutes. Thus, the students were not involved in many P4C sessions.

Moreover, in trials the .programme fidelity is always questionable. One of the schools reported that they stopped doing it, but they were still treated as an intervention school based on the intention to the treat method. Furthermore, it is likely to have a case of John Henry effect (Saretsky, 1972) with the control schools trying to be at the same level with the intervention group. A class of one of the comparison groups reported that they started doing P4C in the middle of the year. This can be considered treatment diffusion.

The comparison group was not impeded from being involved in discussion-based activities. I visited the schools for a day to attend classes (P4C session or not). I

noticed that in students in the comparison group schools were involved in dialogic activities. At the same time, in one of the intervention schools a teacher involved the students in the P4C session without strictly adhering to the P4C guidelines. To be more precise, the teacher guided the dialogue instead of facilitating it. When P4C is adjusted to the school context, teachers may not adhere to its main principles. Hence, its implementation may not differ from any dialogic activity.

This introduces particular bias in the study and threats to the study validity. There is a debate on whether validity should be examined as a unified concept and there is still no consensus (Hughes, 2018). This thesis previously examined facets of validity. At this point, the external and internal validity of the study are examined.

Firstly, the external validity refers to the generalisation of the findings. In the rare cases of random sampling, randomisation can apply to the population that the sampling is drawn (Shadish, Cook & Campbell, 2002, p. 84). The sample of this thesis included English primary schools. However, many schools were invited to the study and refused to participate.

People who chose to participate in a study might have different characteristics from the general population. According to Shadish, Cook and Campbell (2002, p.87) 'when the unit is an aggregate, such as school, the volunteering organisation might be the most progressive, proud, or self-confident' (p.87). The intervention schools participated in the trial of this thesis was already keen on receiving the programme, since they voluntarily signed up for it.

It has to be emphasised that the particular study focused on the P4C impact on critical thinking, creativity and attainment. However, other outcomes, such as the development of other non-cognitive skills, were excluded. Therefore, with the results obtained as a researcher I cannot generalise and recommend that P4C is a positive or a negative programme to be implemented in a school based on the causational relationships this data would establish. I could only claim that P4C has a positive, negative or no effect on the particular domains I examined, but there are other unexamined outcomes which do not allow the generalisation about the overall effectiveness of the programme in general.

Secondly, the internal validity of the study should be discussed. The study should not introduce systematic error (bias) by the design. There are particular sources of bias, which can be threats for the internal validity of the trial as well. According to Torgerson and Torgerson (2008) allocating the participants within the groups in a non-

random way can lead to bias. Indeed in the case of this study, there was no random allocation of the participants within the group, which led to selection bias. For example, it is likely that schools which signed up for P4C training are already more concerned in developing the students' thinking skills and they use techniques for this in everyday practice out of the P4C intervention. In other words, maybe the schools in the intervention group selected in a non-random way might be already focused on the thinking skills development, while the other schools might focus on the development of other skills. This is not a statement to be reported with confidence though. Without external funding offered, it is likely that the schools which consented to participate as comparison group are also the schools interested in the thinking skills reports offered by the researchers. Thus, it is likely that the schools both in the intervention and the comparison group are those with interest in thinking skills.

Another form of bias in the randomised control bias and applicable in this evaluation study is the attrition bias. Not all the questionnaires sent to the schools were returned back and students dropped out before completing the post-test. It is questionable whether the reasons that questionnaires were not returned were the same for both groups. However, attrition and the number of counterfactual cases needed to disturb the finding were reported in the comparative evaluation study of this thesis.

Dilution bias suggests that either people in the control group do not receive the intervention or people in the control group seek for alternatives and they improve their performance based on a different intervention (Torgerson & Torgerson, 2008, p.58). In this study, the schools were sent teacher questionnaires examining how often the P4C sessions were implemented. It might be possible that the teachers did not report the real frequency of the sessions. However, this might be a problem in any self-reported questionnaire. Concerning the comparison group, it would be unethical to demand from the schools to stop trying to develop the thinking skills of their pupils with an alternative intervention. Even though this might introduce bias, it is important to note that P4C in this study was not compared to a comparison group which received no intervention but was compared to a comparison group which received no P4C intervention. The non-random allocation of the participants protects the study from what it is called resentful demoralisation (Torgerson & Torgerson, 2008, p. 60). This occurs when the participants are randomly allocated to a group which does not fit them. In this study, this was not the case. The teachers willing to get a P4C intervention had already contacted SAPERE.

This study did not have an arithmetical balance between the sample in the intervention and the comparison group. This might be considered a disadvantage. However, it has to be mentioned that Gorard (2013, p.128) argued that the two groups do not have to be arithmetically equal. He suggested a limit with the one group being up to three times bigger than the other, with the comparison group usually being bigger as it increases the power with low research cost.

To sum up, there are issues with the external and internal validity of the research. These threats were taken into consideration when the results were interpreted. For instance, based on the consideration of the external validity of the study, there was no attempt for generalisation of the findings.

13.2.1. Measurement Tools

The measurement tools were constructed for the purposes of this research. It is likely that the tools were not sensitive enough to capture a small difference. This study did not find that P4C has an impact on critical thinking, even though a non-standardised assessment was used. Tiruneh, Verburgh and Elen (2014) conducted a meta-analysis to investigate the effective interventions which improve critical thinking in higher education. In their meta-analysis they found that the studies which used non-standardised measurement tools more likely to find that the intervention had a positive impact on critical thinking. This is not the case in the present study of this thesis.

Nevertheless, the same authors also found that the studies which used essays as a measurement tool reported bigger gains on critical thinking compared to those which used multiple-choice questions. This might explain why in this study even though a non-standardised measurement tool was used, no positive gains were reported. Multiple-choice questions are objectively scored. An interpretation of the finding in the meta-analysis about the format of the measurement tools could be based on the expectations of the people who marked the assessments. The marking of essays can be more subjective and the people who marked the critical thinking essays could be influenced by their expectations for the programme to succeed, especially if the marking was not blind. In this present study, blind marking was implemented. However, the expectations of the researcher could not easily influence the results of the study because the scoring of multiple-choice questions is objective.

Furthermore, despite the fact that the tools were piloted, there are indicators of quality of assessments that there were not examined. For example, there is no

Differential Item Functioning (DIF) analysis included in this thesis, which might reveal possible biases for a specific group in the sample, such as specific minorities. In both models, gender was a predictor of performance, so DIF could shed light on this aspect of the assessment.

Even though this might have been an interesting finding, this study did not aim to track identifiable individuals. In order to examine for bias, the researcher should collect demographic data of the students. The collection of demographic data was not considered necessary to answer to this research question and therefore this data was not collected from the participants. As a result, it was not possible to examine whether the assessments systematically biased a particular group of students.

Based on the research question and the consequences of the assessment it was not judged necessary to conduct an extensive analysis on the data for identifying potential bias of the assessments. This is not a problematic decision when the research design is considered, because there was a comparison group. Nevertheless, if in the future these assessments are used to draw conclusions for the individual performance of students, an analysis which can identify potential bias is recommended.

Concerning the measurement of creativity, some of the responses had additional merit (see table 8.4. in Chapter 8). However, the measurement tool of this thesis was not sensitive enough and slightly differentiated responses were still included in the same category. A more sensitive tool might have produced different results.

13.2.2. Creativity

In this section, limitations regarding to the assessments of creativity are reported. When the answers were marked for flexibility and later for prevalence, I had to interpret them and categorise them. This means that the replies were essentially categorised based on my interpretation of the responses. Even though I tried to be as objective as possible, there was subjectivity. This is similar to what Bruner (1999) identified whilst talking about computational science and hermeneutic meaning:

> Say the input into the system is the word cloud. Shall it be taken in its "meteorological" sense, its "mental condition" sense, or in some other way? [...] But to determine which sense is appropriate for a particular context, the computational device would also need a way of encoding and interpreting all contexts in which the word cloud might appear. That would then require the computer to have a look-up list for all possible contexts, a "contexticon"(p.7).

Bruner argued that the same word can have different meanings and the context is what distinguishes the one meaning from the other. Nevertheless, even though the assessments of this thesis were not graded by a computer, there is a common problem. I had no context for each word when I marked the responses since the words were not included in sentences. The students usually named words as uses of objects. In this sense, there is an amount of subjectivity of how I categorised the meaning of the responses of the students. The main indicator used as a context was usually the words that were mentioned before or after a word and similar responses given by other students.

The way I measured prevalence might have disadvantaged students with a literacy problem. Even though general uses such as 'school' and 'making' were accepted as uses of objects, these answers were too generic. Therefore, it was assumed that the most obvious use of these objects was intended. For example, if 'school' was given as a use for a pencil, it was assumed that the students meant writing. Therefore, the prevalence score within the cohort might have disadvantaged students with potential writing difficulties, since the generic answers were generally included and marked as a commonly used category. Another example might be the use of word 'learn'. Even though in the grading of fluency, this word is given one mark as a valid answer and even when combined with 'writing' it is not considered as repetition, in the prevalence scoring, the limits of the interpretation of the response 'learn' required that it is categorised in the same way as 'writing', which is the most common use of a pencil. However, the student might have imagined pencils being used for measurements or another even more imaginative use of pencils for learning.

If the generic answers due to vagueness were considered incredibly creative, that would disadvantage the students who were actually specific. This was not desirable by this research.

There are also limitations concerning the last level of analysis of the creativity activity. The specific number that an answer occurs is sample-dependent. However, I support that the answers which frequently occur in this specific cohort are likely to frequently occur in a different cohort of students of the same age in England.

One of the creativity elements evaluated abstractness of title. Students were asked to give a title for an image they drew. There was an attempt in this activity to exclude measuring the ability of the students to perform well in arts, because this was considered construct-irrelevant when an assessment measures the general creativity

independently of a domain or a subject. There were two main limitations in this activity. First of all, the titles that the students gave were based on the drawing. Therefore, it is likely that the students drew something simple because they did not feel confident about their drawing skill and as a result they might have written a simple title. If this is the case, then the students were led to give a simple title and the abstractness of the title measures indirectly their drawing skills.

The second limitation is related to their literacy. It is likely that the students did not give an abstract title because they lacked the vocabulary to do so. Abstract concepts require higher level of literacy by the students. In order to overcome this limitation students were given the highest marks not only when they provided abstract words in their title, but when they provided titles which were not descriptive. Thus, humorous titles and titles which 'narrated' a story were also given the highest marks in this activity despite the possible lack of abstract vocabulary.

Another limitation is related to the marking process related to creativity. Given that I judged significantly less responses as invalid in the post-test compared to the pre-test, I question the size of negative P4C impact that might have occurred if more invalid answers appeared in the post-test responses. It would be important to replicate this assessment and identify whether the students provide less relevant answers when they were asked about the uses of pencils compared to bricks.

Finally, this thesis measured the creativity as process. This type of measurement has been criticised for having limited predictive validity of future creative achievements and conflicting evidence of content validity by examining only a few aspects of creativity (Said-Metwaly, Van den Noortgate, & Kyndt, 2017). Therefore, if different measurements of creativity could be used in other studies which evaluate the same programme, they could demonstrate an impact on creativity.

13.2.3. Conceptualising Critical Thinking

There is a main limitation in the way critical thinking was conceptualised and assessed. Problem-solving was included in the critical thinking assessment and alternatives were provided because it was evaluated with multiple-choice questions. Nevertheless, it has to be recognised that in real-life the thinkers have to think of the solutions to the problems. Due to this, the predictive validity of the assessment for problem-solving in future tasks and real-life might be restricted.

13.2.4. Regressions

The models created were not exhaustive of all the variables that could have been potentially included. A predictor which was not included in this model was the motivation. Amabile (2017) discussed motivation and the impact that it might have on creativity. There were no variables related to the environment or motivation measured by this thesis. However, their role in creativity development might have been crucial. The reason why only a few variables were included is because only data considered relevant to the research question of the programme effectiveness were collected.

Moreover, even for the existing variables there were not many schools representing each category in the models. For instance, the variable of the years of participating in P4C was included. However, there was only one school which implemented the programme for four years, so it can be questionable to what extent the results of this specific school influenced the models.

13.3. Secondary Data Analysis

At the last section of this chapter, the limitations of the secondary data analysis will be reported. There are some basic limitations in the methodology of this research question. Initially, the separation between the comparison and intervention group is not the most effective. The lists that SAPERE provided included the training places and the date of training. There was no personal information about individuals who received the training. If the training took place in a school, then the school was considered as an intervention school for the analysis. However, it is likely that in some cases some schools hosted the training, but other teachers from local schools attended the training and started implementing P4C in their school. Furthermore, in a few cases the venue of the training was not in a school and SAPERE did not provide information of the people who were registered in the training. For these occasions, there is no information about the teachers and schools which received training. Therefore, they were a few schools which were not included in the intervention group because they received their training in a non-school venue.

Level 1 training for SAPERE does not mean whole-school implementation of P4C. Therefore, the students who had just completed their Key Stage 1 assessments might not have been taught by a teacher who actually implemented P4C on that year or throughout Key Stage 2. However, many schools tend to proceed to whole-school approaches and gain awards for this implementation by SAPERE.

SAPERE did not send lists with trainings events earlier than 2010 and therefore some schools might have received training earlier than this date. These schools were included in the comparison group by this analysis. However, there are two things to be said about these schools. First it was probably unlikely that the school kept implemented P4C and did not receive any additional CPD session by SAPERE after 2010. Since SAPERE has many levels of training for advanced schools, if the school did not receive additional training, it is likely that their practice changed and through time significantly differed from SAPERE training after so many years and therefore they could not have been included in the intervention group. Secondly, if there were schools which received training before 2010 and did not receive additional training or remained in touch with SAPERE might have dropped out from the intervention by this point. To summarise, the schools which received training before 2010 and did not continue to improve their practice were not qualified to be included in the intervention group by the research designed adopted by this thesis.

It was mentioned that the intention-to-treat analysis was followed. Even though some of the schools received training and were included in the intervention group, they might have stopped implementing P4C before Key Stage 2 results in 2016. SAPERE did not have any information about whether the trained schools continued implementing P4C. If the number of the schools which receive training and they stop implementing P4C is high, this will have seriously affected the findings of this thesis. Furthermore, another limitation is that specific teachers sometimes receive SAPERE training. These might change school and start implementing it in a different school. It would be infeasible to call all the schools in England to find out which schools and what classrooms implement P4C. Therefore, it is likely that some teachers moved schools and started implementing P4C in some of the schools in the comparison group or the intervention groups stopped implementing the programme during these four years. For example, trained teachers might have left a school or a new head teacher might have introduced new school targets and priorities.

Even though there was the assumption that the schools which received training before 2012 kept implementing P4C during these four years and were included in the intervention group, it has to be clarified that this is not the case. During the academic year 2015-2016, I randomly selected six schools of the list of the schools which were trained before 2012. I rang the schools in order to find out whether they still implement P4C. I found out that three of them had stopped implementing P4C. Even though I

cannot claim that there is 50% attrition based on such a small sample, when I contacted SAPERE to share my findings, I was informed that such a percent is not very different from the data they hold.

As a result, although the analysis 'intention to treat' was used to answer this research question, I am aware that some schools stopped implementing P4C during these four years. This includes some error in the analysis.

Additionally, SAPERE is not the only organization which provides P4C training to schools in England. Some of the schools in the comparison group might have received P4C training by a different organisation. Despite this being a limitation, it also offered an important advantage to the research design, since there was consistency in the training received by all the schools in the intervention group. I could not examine to what extent the comparison group did not have access to the intervention or similar interventions. It would have been impossible to contact all the schools in the country to find out.

Concerning the analysis, there was a pupil-level analysis. It was assumed that each pupil received or did not receive P4C intervention for 4 years. There was the assumption that pupils remained in the same school during all Key Stage 2. Finally, in the Secondary Data Analysis, a student was considered disadvantaged if they were eligible for FSM the last six years. Even though the eligibility for Free School Meals was used as an indication of disadvantage, this does not cover all the forms of disadvantage. This term is not exhaustive, and a sufficient indicator of all types of disadvantage.

14. Research Findings: Recommendations

So far, there was no study which presented an overall evaluation of the programme. This study summarised and evaluated the existing published evidence regarding the effectiveness of P4C programme. This study gathered evidence which argued both in favour and against the effectiveness of the programme. New evidence was generated with the secondary data analysis and the comparative evaluation study.

In this chapter, recommendations will be made based on all the evidence discussed in chapters 10-12. These recommendations are made for anyone interested in the development of critical thinking and creativity or the P4C implementation. Specific recommendations are made for teachers, school inspectors, P4C trainers, researchers and educational organisations. The recommendations are based on the evidence overall. This means that even though the comparative evaluation study indicated that the programme did not have an impact on students' creativity and critical thinking, I did not disregard the already published evidence which supported the impact of the programme on students' reasoning, literacy and non-cognitive skills.

14.1. Future Research Areas: Literature Gaps

As the systematic literature review made apparent, there are still several gaps to be covered related to the effectiveness of P4C. Despite the fact that there are several studies which examined the effectiveness of the programme, it became apparent that these studies did not always have a strong research design and sufficient reporting. Therefore, their findings were not always judged as trustworthy. More experimental studies - preferably randomised controlled trials - with big sample should be conducted.

This thesis focused on the evidence regarding the programme effectiveness. This evidence only indicates whether the programme worked or not. Although I argued that controlled trials studies have one of the strongest designs to support a causal claim between an intervention and its impact, they do not fully explain why a programme worked or not. Therefore, further investigation of contextual factors is recommended.

The comparative evaluation study of this thesis examined the impact of P4C on several critical thinking skills at the same time and considered a general construct of critical thinking. However, the findings are provisional due to the number of cases who dropped out from the study. Hence, a replication of this study is recommended.

Additionally, there is no examination of the impact of P4C on the learning as a subjective experience. P4C might play a role in motivation longer-term and in fostering positive attitudes towards learning which might be important. Even though the ideas of self-regulating learning and independent learning are promoted by education and P4C, there is no sufficient investigation of the experience of these learners.

Finally, since a new critical thinking assessment was created for the purposes of this study, the psychometric properties of this measurement tool can be examined. The convergent validity of this assessment with a criterion assessment is not available, because there was no other age-appropriate critical thinking assessment. Nevertheless, future research could investigate the divergent validity of the assessment with other assessments of different subjects. I expect some correlation with language assessments and some correlation among the problem-solving items and Maths assessments due to the content of the thinking problems.

14.2. Methodological suggestions for researchers

By conducting this thesis, I realised that conducting a systematic literature review with a calculation of effect sizes, before an empirical study is very important. This identifies the literature gap, but it also creates comparable evidence between the existing research and new evidence. Since effect sizes are calculated, all the evidence is presented on the same scale. Therefore, for the future studies of a similar research design, I recommend demonstrating the literature gap with a systematic review of the literature. Previous studies which used a comparator group can be examined in a systematic way including the calculation of their effect sizes. I argue that approaching the literature in that way increases the value within the thesis because it is not a simple repetition of the evidence, but a creative process. Furthermore, approaching the literature in this way facilitates the comparison with the new findings of the thesis. Also, based on my experience I recommend including the question about the gender as an open-ended question for the participants of this age for the reasons I explained in Chapter 7.

14.3. Recommendations for Teachers

Following the findings of this research, P4C has positive impact on cognitive and non-cognitive skills. Teachers are encouraged to practice P4C in their classroom. P4C is very likely to improve their pupils' reasoning. However, there is no evidence to suggest that critical thinking skills are improved. This might be because of the way P4C is implemented in the classroom. P4C currently emphasises on the justification of

students' opinion (Ventista & Paparoussi, 2016). Therefore, students practice their reasoning and this is reflected on their performance in reasoning assessments. Similarly, teachers should provide their students with the opportunity to practice all critical thinking skills in the dialogue in a Community of Enquiry without restricting them in reasoning.

14.4. Recommendations for P4C Practitioners and Trainers

It is necessary to mention the implications for practitioners. It has already been discussed that SAPERE follows in general Lipman's guidelines in the implementation of P4C. However, during the years Lipman's model has been adjusted. This might be a natural development of the programme, but it might also affect its implementation.

Critical thinking was not found to be improved after P4C sessions. This can raise questions of whether the initial programme used to develop it and - if yes - to what extent the programme that Lipman initially designed has been adjusted effectively. For example, Lipman (1992) recommended specific novels which could both stimulate the dialogue and model the community of enquiry. In the UK, SAPERE trains teachers to be more flexible in their approach and recommend that any material could stimulate the dialogue. According to Chetty (2014), the picturebooks widely recommended by P4C adherents to stimulate discussion on racism failed to address the issue of racism effectively. In his essay, he introduced the idea of a gated community of inquiry where the teachers are not a facilitator but gatekeepers trying to preserve the students by discussing sensitive topics. Hence, it can be questioned whether this flexible approach and the material used to stimulate the P4C dialogue are appropriate and fit the purpose.

Earlier in this thesis, I recommended that the programme should be updated to follow the findings of the new pedagogy and educational research. Hence, I recommend a review of the programme to ensure that the programme even though it does not follow a strict curriculum, it is still able to promote thinking and has clear goals. Similarly, the practitioners should critically evaluate the resources they use to stimulate dialogue in their classroom.

14.5. Time allocated: Recommendations for School Inspectors and Teachers

Currently, even when schools choose to implement P4C, the frequency of implementation varies. For instance, there was a school in the sample of this study,

which chose to implement P4C once a month. The frequency of the programme might play a crucial role for its effectiveness. Siddiqui and Ventista (2018) found that school interventions which aim to improve the non-cognitive skills of the students can show observable impact when they are regularly implemented.

I recommend the programme to be implemented once per week for an hour when the impact of the programme is examined. It is likely that more time is needed for an observable impact to be demonstrated. I think that this justifiable implementation in the school week and therefore teachers and school inspectors should consider this frequency of implementation justifiable. Despite this recommendation, further work needs to be done to examine the optimal frequency of the implementation of the programme in order to have the greatest development of pupils' skills. In the model I created, there were only a few schools in my sample. It is important to examine in closer detail the relation between the frequency and the effectiveness of the programme.

Furthermore, the new Ofsted inspections handbook puts emphasis on the breadth of the curriculum (Ofsted, 2018). According to Spielman (2018), who is Ofted's Chief Inspector, curriculum intents could belong to three different categories; knowledge-led, knowledge-engaged and skills-based. The majority of evidence suggests that P4C develops skills such as reasoning and therefore it can definitely be included in a skills-based curriculum. P4C can also be included in a knowledge-led curriculum because it has impact on attainment. Even though P4C does not involve the teaching of specific knowledge, it has been found that attainment-related skills, such as reading and writing, are developed when P4C is implemented. However, I cannot claim that literacy skills are developed faster during P4C sessions compared to direct teaching of these skills. Ofsted framework might enable the schools to implement this programme, since P4C implementation can be compatible to a rich curriculum.

14.6. Evaluation of Important Educational Outcomes: Recommendation regarding School Funding Allocation

This study was the first study which examined the impact of the P4C programme on critical thinking. There was no available assessment of students' critical thinking skills. This lack of assessments might have caused the lack of evidence of the impact of the programme on these skills. However, if critical thinking is a desirable and valuable outcome of education, then it should be assessed.

Neither the intervention nor the comparison group developed their performance in the critical thinking assessment. The assessment I created might not have been effective to capture differences in the performance of the two groups. This skill might require more time to develop. It should be questioned when impact should be expected for these skills. However, it might also mean that no change was made. If within the year, the comparison group did not improve their performance, then it can be questionable whether education has an impact on this skill. It is likely that education fails to develop critical thinking.

Thus, a recommendation for the future should be made about the evaluation of critical thinking. If critical thinking is a valuable outcome of education, more assessments should be developed. These methods should be triangulated to demonstrate whether the students improved their critical thinking. Even though teachers might report improvement in the critical thinking skills of their pupils, this is not a sufficient indicator of a real change in critical thinking. The number of critical thinking assessments is limited. In many cases, schools have to buy assessments and these assessments may not be affordable by schools. Assessments measuring valuable outcomes for education should be freely available to schools.

14.7. Closing the attainment gap: Recommendations for policy makers

All the research questions of this thesis were summarised as one main question. By gathering all the available evidence should somebody implement P4C in the school? This thesis attempts to answer this question to support anyone who would like to implement an evidence-based policy.

The main recommendations for the educational policy are discussed in-depth in the next chapter. However, the main recommendation that can be made based on the findings of this thesis is that people who decide on the educational policy should include P4C in the curriculum if they aim to close the potential attainment gap between advantaged and disadvantaged students. Various studies demonstrate that the programme has gains on disadvantaged students' skills and attainment.

14.8. Recommendations for Evidence Based Educational Organisations

Since P4C should be included in an evidence-based policy, this section includes recommendations for organisations which support the implementation of Evidence Based Policy. For example, Education Endowment Foundation (EEF) funds research in

order to create evidence for education and creates resources to support evidence based education. Moreover, initiatives such as ResearchED claim to help to link this evidence with the school practice by organising events for teachers. As it was discussed in Chapter 12, this study had similar findings to previous research funded by EEF.

As the table of the systematic literature review also included the context of the studies, such as the age of the participants and country of implementation, these organisations should examine and report the context of the interventions to see whether they judge that similar results will occur in their context. As a result, as a methodological suggestion, organisations such as the EEF, should also report the context of the interventions when they summarise the evidence.

EEF toolkits currently report the security of the available findings. Based on the number of cased needed to disturb the findings, the findings of my study are not secure. When an intervention is implemented, this means that time is not given to a different intervention. My study did not include any comparison to other interventions to be able to suggest whether time and resources should be allocated to P4C compared to a different intervention.

Despite the fact that there may be interventions which report a bigger impact on these domains, this study also suggests that P4C implementation is justifiable. EEF reporting is mainly focused on attainment. Siddiqui and Ventista (2018) discussed the importance of developing the non-cognitive skills and promising school-based interventions. P4C programme was found to have a positive impact on different cognitive and non-cognitive skills.

14.9. Recommendations for Teacher Education

The comparative evaluation study of this thesis showed that P4C did not have an impact on students' critical thinking and creativity. This finding made weaker the overall consistently picture which showed that P4C develops students' reasoning skills. However, there were robust studies which supported the impact of the programme on reasoning. Given that all the previously published studies consistently found the positive impact of the programme on reasoning and most of the times a positive impact on literacy and some non-cognitive skills, I suggest that teacher education programmes should train the prospective teachers to implement P4C. P4C should be encouraged by programmes which train teachers for liberal education. As Demissie (2017) suggested P4C can support trainee teachers in their reflective practice. Similarly, I argue that it

can train them to teach a skills-based curriculum. Trainee teachers can practice being a facilitator instead of an authority in the classroom. I encourage teacher training programmes to include P4C training.

14.10. Implications for the nature of thinking skills

This thesis examined the critical and creative thinking. There are authors, which discussed the relationship between these two thinking skills. For example, Fisher (2010) introduced the term *critico-creative thinking*. This thesis provided evidence against this theoretical position. Critical and creative thinking were examined separately. Critical thinking was examined as purposeful, reflective thinking, whilst creative thinking was perceived as divergent thinking. The evidence suggested that these two skills might be taught and developed with different types of activities.

Existing evidence suggested that P4C leads to the development of reasoning. On the contrary, there was no strong evidence to suggest that it leads to the development of divergent thinking. This thesis examined the impact of P4C on divergent thinking and found a negative impact. This suggests that an intervention which might consistently develop reasoning, it might not develop divergent thinking. Therefore, critical and creative thinking are not linked to the extent that some bibliography claims.

This finding can have implications for the development of these skills in schools. Since these two types of thinking differ, the same intervention should not be expected to develop both types of thinking. Reflective thinking is purposeful and moves to a specific direction, whilst divergent produces different ideas which might have different directions (fluency). Different activities require purposeful thinking. However, this thesis also questioned whether divergent thinking is a sufficient indicator to measure creativity.

14.11. Creativity Findings: Implications for Workplaces

When creativity results were analysed, an interesting finding related to the nature of creativity occurred. The students who provided more responses were not always those who provided the most innovative responses (see Tables 11.14 and 11.15 in Chapter 11). This can have implications related to the nature of creativity and its expressions in everyday life.

Related to the nature of creativity, it suggests that innovative responses are not always produced by people who score highly in divergent thinking tests. In that sense,

creativity as innovation is only deviant without always being divergent. Therefore, this suggests that the standard definition of creativity (Runco & Jaeger, 2012) has two sufficient criteria. According to the standard definition, creativity requires originality and effectiveness. This definition does not involve the quantity of production. From the results of this thesis, it became apparent that people who produce more responses do not necessarily produce more innovative responses. This might mean that divergent thinking, which includes fluency and flexibility, is not always a good indicator to assess and judge which person is the most creative.

Some creative people produce less in quantity, but they are still able to generate high-quality and innovative products. This can have implications on how creativity should be measured in the future and how the creative person should be perceived. For example, if productivity in school or workplace is currently measured with the quantity of production, the findings of this study demonstrate that this is a misleading indicator about the innovation of the students or employees.

This finding can have applicability to different contexts, such as higher education or businesses. For example, in higher education the researchers and academics are expected to produce many research outputs. However, according to my findings it can be argued that these people who produce the more outputs are not always the same people who produce the most innovative research. This finding might also suggest that people who allocate more of their time thinking, they can produce more innovative answers compared to these who use the same time to produce more responses. In some cases, the quantity of production compromises the quality of creative responses.

14.12. Implications for Assessing Creativity in P4C sessions

This thesis showed that P4C sessions had a negative impact on students' divergent thinking. If creativity is accepted as concept which can be taught and assessed, the finding of this study implies that P4C is ineffective for the development of creativity. However, creativity might be taught with explicit teaching and P4C currently does not seem to have an explicit teaching or usual practice of divergent thinking.

Based on the findings of this thesis, I think that future studies, which will attempt to examine the impact of the programme on creativity, should not examine its impact on divergent thinking. This is only an aspect of creativity and maybe a narrow one. In Chapter 3, different aspects of creativity were presented. Torrance (1988)

included problem-finding in the creative process whilst Piirto (2010) argued that the creative person is curious. These may be two aspects which are currently practiced in the community of enquiry instead of divergent thinking. The idea of community of enquiry is based on students' natural curiosity. In the community of enquiry, students question. These were possibly creative elements developed in P4C sessions, but they were not examined by this thesis.

Since I have already argued that purposeful and divergent thinking are different types of thinking, this might suggest that different type of activities develop critical and creative thinking. However, there are elements of creativity linked to critical thinking. Therefore, these aspects might be developed in P4C sessions. I believe that if P4C does not include any activities which target the development of divergent thinking, this is not the type of creativity to be assessed after P4C sessions. Future researchers and teachers should assess the creativity linked to critical thinking, creativity as process linked to problem-finding, question generation and curiosity as a trait. These are the creative elements currently practiced in P4C sessions and there is no robust evidence to suggest that P4C has or does not have an impact on these.

15. Conclusions

So far, this thesis presented and discussed the evidence on the effectiveness of P4C and its impact on the cognitive and non-cognitive skills of primary school pupils, especially those from disadvantaged backgrounds. P4C is a skills-based intervention since it does not teach students' specific knowledge but aims to improve a range of skills. This final chapter will summarise the evidence about the impact of the programme on different cognitive and non-cognitive skills. Specifically, this chapter responds to five questions:

a) Does the programme improve students' cognitive and non-cognitive skills?
b) Should P4C be implemented in primary schools in England?
c) If yes, how should P4C be implemented?
d) Is attainment developed by a skills-based intervention?
e) How can schooling support students' thinking skills?

15.1. Does the programme improve students' cognitive and non-cognitive skills?

This thesis responds to the question of whether the implementation of P4C is evidence-led, and whether P4C is an effective intervention for the improvement of students' cognitive and non-cognitive skills. The thesis reviewed the evidence on the programme effectiveness published in the last forty years, almost since the programme was founded, in order to provide a clear aggregated answer about programme effectiveness. This evidence did not draw a consistent portrait of programme effectiveness, as it might be expected because P4C is an intervention implemented in widely different educational contexts. The thesis then produced new evidence in areas that required further investigation, by conducting a comparative evaluation study and investigating attainment data available from the Department for Education, to provide a holistic evaluation of the programme.

Overall, there is evidence from the review and this new study that the programme is likely to have a positive impact on attainment and close the attainment gap between advantaged and disadvantaged students. The secondary data analysis showed that students eligible for Free School Meals develop their reading and writing at Key Stage 2 after long-term P4C implementation compared to non-P4C practice.

This thesis widely discussed the importance of developing the thinking skills of pupils in primary schools. P4C sessions are not focused on a particular knowledge and they are skill-based. Lipman (2003), who founded the programme, argued in favour of

an impact on critical, creative and caring thinking of students. Therefore, this thesis examined the extent to which thinking as a skill can be improved by these sessions. When creativity was concerned, the programme had a slightly negative impact on divergent thinking. Similarly, for an overall development of critical thinking skills, P4C was not found to be effective. Even though the results of previous studies suggest that P4C has an impact on students' reasoning skills, the comparative evaluation study presented in this thesis found no sound evidence to suggest that P4C was beneficial for students' critical thinking or creativity skills.

There is adequate available evidence for the positive impact of the programme on reasoning skills, as shown by the many studies examined this area. However, the findings of this new comparative evaluation study did not support this consensus. Of course, the findings of one study are not sufficient to support or deny the effectiveness of ineffectiveness of the programme. Except for this comparative evaluation study, all the previously published evidence suggested that the programme has a positive impact on students' reasoning skills. Different sizes of positive impact were reported and this might vary based on the context and the measurement tools adopted in the studies. However, studies conducted in different contexts reported a positive impact on reasoning. Furthermore, this positive impact is retained as long-term studies and studies with a follow-up suggest. My evaluation study made the consistent picture regarding the effectiveness of the programme on reasoning weaker. However, it did not overrule these findings. By combining the available evidence, it can still be suggested that the programme is likely to have a positive impact on students' reasoning skills, at least in a small way.

P4C could have an indirect impact on students' thinking skills. There is evidence form robust studies that P4C does improve the literacy skills of students to some extent (compared to not having P4C). Furthermore, this thesis found some evidence that students who participates in P4C sessions are more likely to use abstract concepts. Therefore, P4C might lead to vocabulary enrichment. The secondary data analysis of this thesis showed that it improves the literacy skills of disadvantaged students, in terms of standardized attainment. If P4C improves the literacy skills of students, it can be argued that it also improves their thinking. Dewey (1933, p.230) argued that even though language is not thought it is necessary for thinking and communication. If there is a relationship between language and thinking and P4C improves the first, then it could have an impact on the latter.

There was also published evidence concerning the impact of the programme on non-cognitive outcomes. Students who participated in P4C sessions reduced their mean scores on disappointment, grandiosity and instability and increased their scores in meta-cognition, adaptability and determination measurement tools.

This is an interesting finding, particularly when non-cognitive issues related to childhood are discussed. For example, the latest published report by Office for National Statistics focuses on loneliness, and revealed that 14% of children aged 10-12 years old report feel lonely and 27% of children 10-15 years old eligible for Free School Meals often feel lonely (Snape et al., 2018). This suggests that schooling should not only be concerned with the cognitive skills of students and their attainment. Schooling should develop students' non-cognitive skills as well. P4C is found to develop non-cognitive skills, and therefore it can indirectly support students to deal with life-related issues. Even though there were mixed findings of the programme on its effectiveness on improving students' co-operative skills, it might still provide solution to problems such as the current loneliness in the young population. As a dialogic intervention, P4C is suitable to increase the dialogue and communication between the students and to encourage students to interact with each other.

Concerning attainment, it can be expected that the programme is likely to have a more positive effect in areas such as reading, writing, use of vocabulary involving abstract concepts and meaning development. Overall, it can be argued that the programme supports students' thinking skills directly by developing their reasoning skills and indirectly by increasing their literacy. However, the thinking that P4C improves is the purposeful, reflective thinking instead of divergent thinking.

15.2. Should Philosophy for Children be implemented in schools in England?

Currently, the Education Endowment Foundation in England includes this programme as one of the promising projects for developing effective learners (Education Endowment Foundation, 2018b). However, results from a single project are probably not adequate to support an evidence-based policy. Hence, this thesis questioned whether P4C improves attainment and wider outcomes in education. This intervention develops some cognitive and non-cognitive skills. P4C currently is only found to have a negative impact on students' divergent thinking. However, the benefits of the

programme outweigh the disadvantages overall. Therefore, schools in England can include this intervention in their curriculum.

The comparative evaluation study of this thesis was the first large-scale evaluation study which examined the impact of the programme on critical thinking skills and creativity. The study found that P4C has no impact on Year 5 students' critical thinking skills and it has a negative impact on their fluency and innovation. There is no comparable evidence to support or contradict these findings. However, several studies showed that P4C has a positive impact on reasoning skills. As a result, primary schools can implement this intervention to develop students' reasoning skills.

The programme was found to have a positive impact on reading and writing of students eligible for Free School Meals. The reason why FSM students develop their attainment might be that P4C sessions do not require a 'right' answer. Thus, students who might not perform well in other subjects have the opportunity to express their opinion. This might improve their language skills since they practice their speech and oral literacy by participating in the dialogue. I particularly recommend this programme for educational policies which aim to close the attainment gap between advantaged and disadvantaged students. This is because the disadvantaged students consistently seem to benefit when they are involved in P4C session and this does not only refer to their attainment.

Nevertheless, if the aim is the improvement of Maths ability, this programme is not recommended. Despite some evidence which showed a small impact of P4C implementation on maths ability (Gorard, Siddiqui & See, 2015), there is no additional evidence to support this claim and it was not confirmed by the long-term analysis of this thesis. Maths ability can only broadly be associated with P4C and thus P4C had a slightly negative impact on it. Hence, not participating in this programme can actually provide more opportunities and time to improve students' mathematics ability.

To summarise, based on the priorities of the educational policy, P4C can be included in the educational agenda. If closing the literacy attainment gap between advantaged and disadvantaged students is prioritised, then P4C is a school-based intervention which should be implemented. Similarly, if the schools or individual teachers prioritise reasoning skills, non-cognitive skills and language development, schools include P4C in their curriculum.

15.3. How should P4C be implemented?

P4C should be implemented when teachers, school leaders or policy makers would like to support disadvantaged students in the classroom since the evidence suggests that the programme is more effective for these students. Since P4C dialogue does not have a right or wrong answer, it creates a supportive environment for students who do not perform highly in other subjects to express their opinions and participate.

In order for this environment to be created and for the students to have this 'space' and time to express themselves freely, P4C should be assigned time slots which promote this liberal education where no specifically pre-requisite knowledge background is required for active participation. Specifically dedicated time in the schedule of the students is needed and P4C should not be imbedded in different subjects in order to be implemented properly.

It is important to emphasise that P4C should not be implemented in the time of religious studies, which is something that some schools might do as noted in other trials (Gorard, Siddiqui & See, 2015). For example, by observing a school in the trial I noticed that the topics discussed in P4C session seemed to be religious related. The teacher discussed the topic of forgiveness and whether students should always forgive the others and in which cases. Despite the fact that P4C does not aim to reach one right conclusion, the teacher was not simple a facilitator and aimed to demonstrate the importance of forgiveness and demonstrate that forgiveness is beneficial for the person who forgives. This is an additional reason that P4C should have specifically dedicated time in the weekly timetable and should not replace existing subjects.

P4C is usually implemented once a week. This seems like a reasonable minimum amount of time for the school-intervention to be implemented in order to lead to the development of some cognitive or non-cognitive skills of students.

Concerning the implementation of the programme, fidelity is crucial. SAPERE provides training to teachers. Teachers adjust it when applied in their context. However, it can be questionable to what extent this adjustment is justifiable and to what extent this implementation loses the basic elements of the initial intervention. The example of the teacher embedding the session in a religious-related session clearly demonstrated that when a programme is included in the classroom, its form might change and it might be the case that this programme is not the initial programme. In this case, the teacher wanted the students to reach a specific conclusion which is against the P4C nature, where questioning is more important than the answers

(Ventista & Paparoussi, 2016). Hence, despite the adjustments, P4C trainers and teachers should clearly set and understand the P4C principles in order not to seriously compromise its fidelity. At the same time, as it was argued in Chapter 2, P4C is a pedagogical intervention and should follow the pedagogical updates.

To summarise, the way of implementation should enable adjustments according to the context. However, P4C should not lose its democratic character. For example, no matter how the students get permission to speak with raising hands or not, in no case it will ever be acceptable not to allow students to speak freely. Similarly, P4C should facilitate students' questioning. Losing this element of the session is an unacceptable adjustment.

15.4. Is attainment developed by a skills-based intervention?

Hirsch (2001) is an example of scholar who argued in favour of a knowledge-based curriculum. He supported that general knowledge highly correlates to general ability to learn and the ability to learn is domain-specific. P4C aims to equip the students to evaluate the given knowledge, but currently there is no evidence of whether P4C can succeed to this aim. Hence, it can be questionable whether school-based interventions focused on knowledge should be implemented.

However, thinking skills might not be malleable as easily as literacy. If P4C is not effective in improving the thinking skills of the students, it could be argued that it would have been better to implement a knowledge-based curriculum to – at least- support their attainment. Recent study in England, however, demonstrated that general knowledge cannot support literacy and there is only weak evidence that Free School Meal students' literacy can be improved using a knowledge-based curriculum over a short period of time (See, Gorard & Siddiqui, 2017). Therefore, even though P4C was not initially designed mainly to improve literacy, it supports students' literacy more than core curriculum which is based on knowledge. Hence, it cannot be claimed that P4C which is a domain-independent and skill-based intervention takes away time needed for the core curriculum because P4C has a positive impact on literacy.

This thesis showed that there is evidence from a randomised controlled trial which suggest that P4C as a subject-independent dialogic intervention developed the attainment scores of students. These effect sizes were higher than those reported in the trial with the knowledge-based curriculum. This finding refers to literacy skills. This might be due to the fact that P4C requires from the students to practice literacy skills.

Therefore, these skills can be developed independently of a context. A skills-based curriculum is sufficient to develop students' literacy. This verifies Paul's argument mentioned in Chapter 3. Writing and reading are general abilities. Even though there is writing about X or reading about X, it is possible for the students to learn and write or read in general.

15.5. How can schooling support students' thinking skills?

In Chapter 3, this thesis presented some evidence to support that critical thinking and creativity as general skills can be developed. This evidence might be questioned for their robustness. As I also argued in Chapter 3, even if there is no known intervention to improve these skills, educational research should investigate ways to develop them because of their importance.

According to the comparative evaluation study conducted and discussed in this thesis, P4C did not have an impact on critical thinking or creativity. The intervention group did perform better than the comparison group in problems which require basic critical thinking, such as inference and evaluating the credibility of the sources. An explanation for this might be the fact that the programme stopped following Lipman's curriculum. Whilst the characters in the novel of Lipman facilitated modelling of thinking, today the material used is selected by the teachers and it is less strictly selected. This does not necessarily lead to the lack of improvement of thinking skills. If the programme returns closer to what Lipman suggested, thinking skills might improve. However, there is no published evidence to suggest this. It is evident though that P4C implementation is currently less structured.

The finding of this study can be linked to the existing evidence about the effectiveness of cognitive training on general cognitive ability. General cognitive ability which is commonly referred as intelligence is domain-independent and has been associated with different tasks. Evidence suggests that this ability is not improved when the students get training in specific cognitive tasks (Sala & Gobet, 2019). For example, activities which involve various cognitive skills, such as playing video games, chess or music instruction, do not lead to the development of general cognitive ability. On the contrary, the researchers suggested that interventions boost performance only on tasks similar to the trained task.

I argue that this finding might suggest that training in specific tasks does not improve general ability. It might be difficult for the students to transfer these skills in

different contexts. In order for the students to transfer these skills, they should be able to generalise these skills, form general principles and transfer them in new contexts. However, as I argued in the theoretical chapter, students do not have to be able to formulate new principles in order to be critical thinkers. This is a very demanding task. If the transferability of the skill requires abstraction and formulation of general principles, the students might fail to generalise these skills.

The comparative evaluation study of this thesis also showed that an intervention which accepts critical thinking and creativity as general abilities, but does not teach them explicitly cannot improve the thinking skills. P4C is a dialogic intervention in a community and therefore the students argue, question, express and justify their opinion during the sessions. According to this element, reasoning, non-cognitive skills and language skills (literacy) used during the participation of sessions, whilst Maths, creativity as process - as defined by this thesis- and other critical thinking skills are not. The programme expectedly leads to the development of only the initially mentioned skills because these are the skills being practiced during the sessions.

Therefore, I argue that if teachers aim to develop general thinking abilities of their students, they should provide them general training which can be applied to different contexts. This will make the transferability of the skills easier. Students should learn criteria and techniques which can be applied in different contexts. Training which accepts these abilities as domain-independent can provide general principles to students and give them the opportunity to practice them in different contexts.

If critical thinking and creativity are improved only by explicit teaching and practice of the skills, then P4C should work towards this area. Published evidence (not the comparative evaluation study of this thesis) suggests that P4C currently improves reasoning and this might be because it engages students in discussion when they practice their reasoning. Hence, the findings about the effectiveness of the programme in relation to specific skills can be interpreted when the elements of P4C programme are examined.

A school-based intervention can develop critical thinking and creativity as general and subject-independent skills. In order to achieve this, students should practice these skills. Skills-based curricula which encourage the development of

transferable and subject-independent thinking skills should facilitate the explicit teaching and practice of these skills in different contexts.

15.6. Concluding Thoughts

To summarise, some of the ideas discussed in this thesis are presented below:

- P4C can be a valuable intervention. Teachers need training in order to implement this programme. The training of teachers can take place either during their initial teacher education or as a part of their Continuous Professional Development.
- Based on the evidence overall, P4C is found to develop students' reasoning skills in different contexts. This impact may be retained if the students are followed-up after the end of the intervention.
- On balance, P4C can support different skills of disadvantaged students and can contribute closing the attainment gap between advantaged and disadvantaged students. The available findings suggest two ways of closing the attainment gap. When P4C intervention was implemented for one academic year, there was an increase of the literacy skills of both advantaged and disadvantaged students. The disadvantaged students had bigger gains compared to the advantaged ones. However, the secondary data analysis showed that long-term implementation of the programme led to an increase of the scores only of students eligible for FSM and a decrease for the rest of group.
- The educational programmes require adjustments. These adjustments should mainly correspond to the prevailing pedagogy. Some adjustments to the school context are necessary. However, the school-based interventions should retain their main elements. This recommendation should apply to P4C as well.
- Skills, such as literacy and thinking, can be general and transferable to different contexts. Even though knowledge is subject-specific, specific knowledge is not required for the development of the general skills. This study demonstrated that there is evidence which suggest that in many studies P4C as a dialogic skills-based intervention developed students' literacy.
- There may be a contradiction between the development of reflective (or critical thinking) and divergent thinking. Reflective thinking is purposeful thinking moving towards a specific direction, whilst divergent thinking and specifically

flexibility operates differently. Ennis (1985, 2015a) argued that critical thinking is 'focused' and has a specific purpose. This seems to be the opposite of what divergent thinking suggests even when presented as 'functional creativity'. P4C is often found to have an impact on students' reasoning. However, the comparative evaluation study found secure evidence that it had a negative impact on students' divergent thinking. It can be questioned whether these two types of thinking use the same mechanisms and how the same dialogic intervention can develop both. It can be questioned to what extent the same school-based intervention can develop both types of thinking. I argue that different types of activities develop these two types of thinking.

- Previous studies show that explicit teaching of critical thinking can develop critical thinking and creativity. P4C is a skills-based intervention which does not explicitly teach thinking skills. Based on the finding of the comparative evaluation study with Year 5 students, students' thinking skills were not developed with implicit teaching.
- There was no large-scale evaluation study examining the impact of the programme of creativity. For the first time, the comparative evaluation study presented in this thesis examined the impact of the programme on Year 5 students' creativity. Evidence showed that the programme has negative impact on students' divergent thinking. This evidence was judged secure based on the fact that the number of counterfactual cases needed to disturb the finding was smaller than the attrition. However, it is likely that the programme has a positive impact on different elements of creativity, which were not measured in this thesis. For the future studies, creative skills explicitly practiced in P4C sessions should be examined. The impact on curiosity as a trait of the creative people and on problem-finding as an element included in the creative process should be investigated.

Appendix

Appendix 1. Chapter 5.

Appendix 1a. Telephone Guide for the participation of the schools in the project.

Good morning/afternoon. My name is Ourania Ventista from Durham University and I am calling regarding a research project for philosophy for Children. Could I please talk to…(*Ask the Contact Name of the Specific School)?*

waiting time

Hello. I am Ourania Ventista from Durham University. I e-mailed you a few days ago concerning the research for philosophy for children. I am calling because I would like to invite your school to participate. Do you have a couple of minutes to be informed about it or would you like me to call a different time?

Hopefully they will say yes – if not I can ask when I can call back

Thank you! Well, your school is invited to participate in a research. This research examines the P4C impact on creativity and critical thinking and it is only for Year 5 students. Does your school administer P4C in Year 5 classroom?

Response

Great! More than 30 schools who implement Philosophy for Children will be invited to participate. The participation is mainly a really short assessment of creativity and critical thinking. It lasts only 20 minutes and will be implemented twice once in September 2016 and once at the end of the same school year (May or June). You will receive the survey questionnaires by post and the only thing that it will be demanded from the school is to have a supervisor during the survey to read the instruction and supervise and at the end to post it with a pre-paid envelope. After the two surveys we will provide you with the results of your school and the overall results. That was a brief description. If you are interested I could forward, you an e-mail with more information to your e-mail address. Does this sound ok to you? Do you have any questions regarding this?

I will write down the questions.

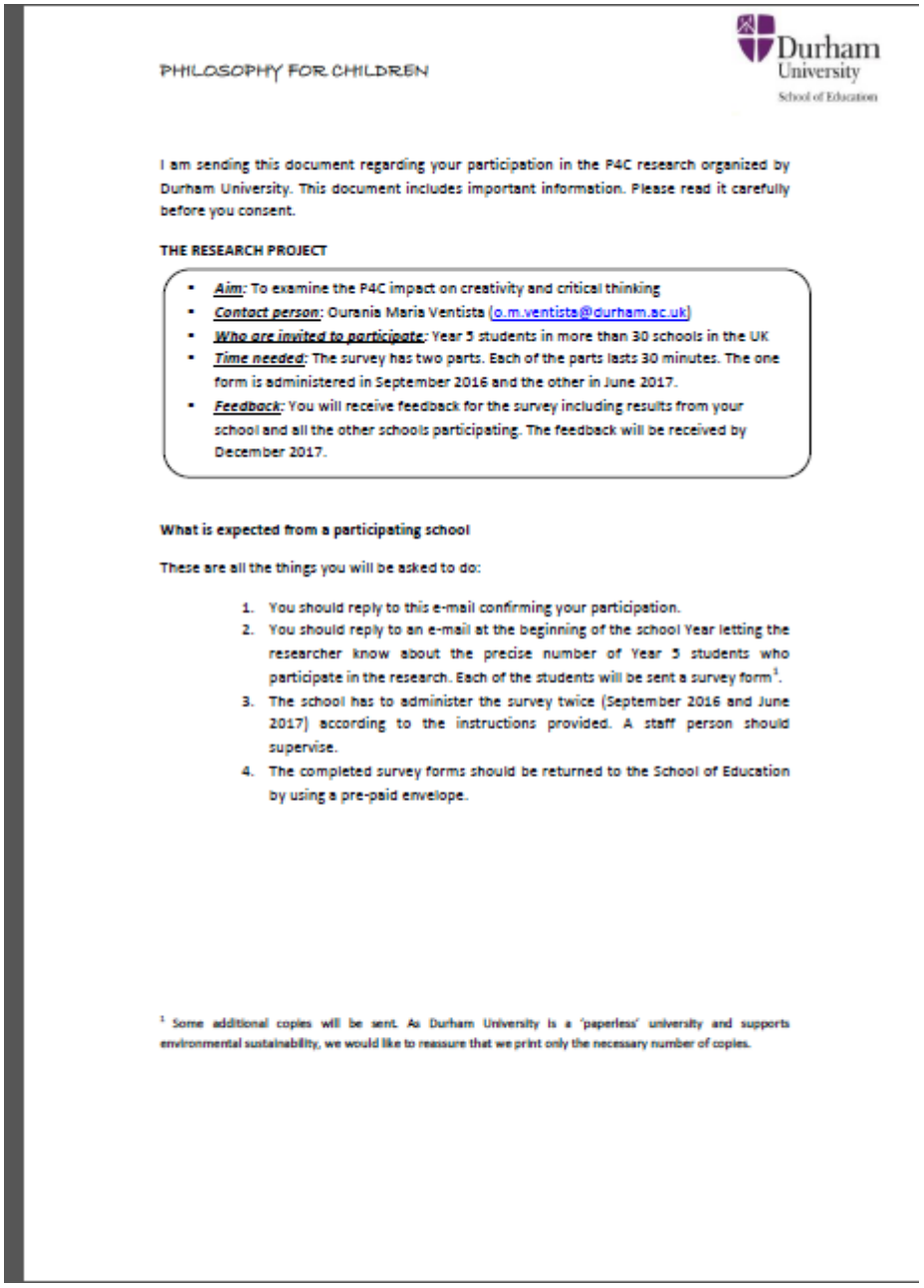
These are interesting questions. I will forward you the e-mail and I am confident that the information provided will give an answer to your questions. In case you have further queries, or you want more information, you can of course reply to my e-mail and I will get back to you as soon as possible. Is this ok for you?

In the same e-mail I will also attach a participation sheet which should be completed by the head teacher of the school in order for any school to be a part of the research. But if you have any questions, please feel free to respond to my e-mail before you sign.

Give time to the person to talk! – if it needs take notes

Thank you very much for our telephone conversation. It was nice talking to you. So, in the next 24 hours you are going to receive an e-mail from me with the information and the participation form. We will be glad to hear from you and have the school participating in the research.

Appendix 1b. Information pack emailed to the schools (after the telephone conversation)



PHILOSOPHY FOR CHILDREN

Durham University
School of Education

I am sending this document regarding your participation in the P4C research organized by Durham University. This document includes important information. Please read it carefully before you consent.

THE RESEARCH PROJECT

- ***Aim:*** To examine the P4C impact on creativity and critical thinking
- ***Contact person:*** Ourania Maria Ventista (o.m.ventista@durham.ac.uk)
- ***Who are invited to participate:*** Year 5 students in more than 30 schools in the UK
- ***Time needed:*** The survey has two parts. Each of the parts lasts 30 minutes. The one form is administered in September 2016 and the other in June 2017.
- ***Feedback:*** You will receive feedback for the survey including results from your school and all the other schools participating. The feedback will be received by December 2017.

What is expected from a participating school

These are all the things you will be asked to do:

1. You should reply to this e-mail confirming your participation.
2. You should reply to an e-mail at the beginning of the school Year letting the researcher know about the precise number of Year 5 students who participate in the research. Each of the students will be sent a survey form[1].
3. The school has to administer the survey twice (September 2016 and June 2017) according to the instructions provided. A staff person should supervise.
4. The completed survey forms should be returned to the School of Education by using a pre-paid envelope.

[1] Some additional copies will be sent. As Durham University is a 'paperless' university and supports environmental sustainability, we would like to reassure that we print only the necessary number of copies.

PHILOSOPHY FOR CHILDREN

Durham University
School of Education

FREQUENTLY ASKED QUESTIONS:

Do the students complete the form anonymously or are they asked to write their names on the forms?

The research examines the P4C impact on critical thinking and creativity. This can be examined without tracking individual performance. However, if the school wants to track individual performance this feedback can be provided. Therefore, it is the school's choice whether it wants individual or classroom feedback. In case you want individual feedback, then the students can complete their names on the forms sent. If you want to track the progress of the classroom in total and you are not interested in the individual performance, then the students will not be asked to complete their names on the forms. This decision is only about the feedback that the school wants to receive. For the research project we will analyse the anonymous classroom progress.

What type of questions will be included in the forms?

The questions will be activities which assess some aspects of creativity and critical thinking. No preparation is demanded to reply them. The majority of the questions will be multiple-choice. For critical thinking section, the questions will be simple problems. The students should answer reasonably mostly by using their 'common sense'. For creativity, the students will be expected to provide innovative responses.

Appendix 1c. Ethics Approval Letter

Durham University
School of Education

Shaped by the past, creating the future

2 March 2016

Ourania Maria Ventista
PhD

o.m.ventista@durham.ac.uk

Dear Ourania

What are the longer-term and high-order thinking outcomes of Philosophy for Children in Primary Schools?

I am pleased to inform you that your application for ethical approval for the above research has been approved by the School of Education Ethics Committee. May we take this opportunity to wish you good luck with your research.

P. M. Holmes

Dr. P. Holmes
Chair of School of Education Ethics Committee

Leazes Road
Durham, DH1 1TA
Telephone +44 0191 334 8300 Fax +44 0191 334 8311
www.durham.ac.uk/education

Appendix 2. Chapter 7.

Appendix 2a. The letter included in the envelopes sent to the schools

Dear Teacher,

Thank you for helping with this survey. In the envelope you received, you can find:

a) An administration guide. The administration guide will give you guidelines to follow for the administration of the survey.

b) The survey forms.

c) A form to be completed after the survey by you (or the person who helped with the survey).

Should you have any queries, please do not hesitate to contact me. My e-mail is o.m.ventista@durham.ac.uk. My phone number is 07918519506.

Yours sincerely,

Ourania Maria Ventista

School of Education
Durham University

Appendix 2b. Administration guides for the pre-test and post-test

ADMINISTRATION GUIDE

The survey should be completed by Year 5 students and on any day between 15th September and 15th October. The students should not be rehearsed for this in advance. The administration lasts no more than 30 minutes. On the day please follow the following steps. The underlined sections are instructions which should be read to the students.

STEP 1. Give one form to each student and please read the instructions to the students before they start writing.

INSTRUCTIONS

- You have 30 minutes to complete these forms.
- Please listen to the instructions carefully and do not go on to the next page unless told to do so.
- On the top left part of some of the following pages, you will find a cloud where the people who made the tests wrote give some help. Please read the help before you continue.
- Please do not spend more than 10 minutes on the first two activities 'Uses of Object' and 'Drawing'. Otherwise, you will not have enough time for the thinking problems.
- If you have any questions, please ask now.
- Go on to the next page.

STEP 2. After the first 10 minutes of the test, please advise the students to move to the thinking problems.

- For the thinking problems you have to choose one of the options given and write a letter (A, B or C) in the box provided. There is **only one best answer** for each of the thinking problems.
- Do not guess. If you cannot decide, leave it blank and move to the next question, because you have only a few minutes to complete the questions.
- At the end, if you have time left you are allowed to revise your answers.
- When you finish the test, please remain seated. When the 30 minutes are up, you will be informed.

STEP 3. After 30 minutes, please inform the students that their time is up and request that they put their pencils down. Ask a student to gather the completed tests, while you are ensuring that nobody takes more time to write.

STEP 4. Please complete the post-administration form.

STEP 5. Use the pre-paid envelope to post the surveys back to the School of Education.

Should you have any queries, please do not hesitate to contact me. My e-mail is o.m.ventista@durham.ac.uk. My phone number is 07918519506.

Frequently Asked Question:

What help am I allowed to provide to my students?

During the administration, you can help your students to understand the questions (e.g. you can read the questions to a student with low reading ability or visual impairments) but you should not provide them with hints to respond to the questions. Please do not suggest the correct answer. Even though you can offer help with reading, please do not paraphrase the questions. The vocabulary used has been piloted and confirmed as age appropriate.

What should I do if a student is absent?

The absent students can be given the opportunity to complete the survey form on a different day or at a different time. However, please keep a note on the top of the form that the specific student was absent and please try to ensure similar administration procedures for the student.

ADMINISTRATION GUIDE

The survey should be completed by Year 5 students on any day between 12th May and 12th June. The students should not be rehearsed for this in advance. The administration lasts no more than 30 minutes. On the day please follow the following steps. The underlined sections are instructions which should be read to the students.

STEP 1. Give one form to each student. There are two versions of the test. Please give the

VERSION A to the students who completed the survey at the beginning of the school year.

VERSION B to the students who did not complete the survey at the beginning of the school year.

STEP 2. Please read the instructions to the students before they start writing.

INSTRUCTIONS

- You have 30 minutes to complete these forms.
- Please listen to the instructions carefully and do not go on to the next page unless told to do so.
- On the top left of some of the following pages, you will find a cloud where the people who made the tests offer some help. Please read the help before you continue.
- Please do not spend more than 10 minutes on the first two activities 'Uses of Object' and 'Drawing'. Otherwise, you will not have enough time for the thinking problems.
- If you have any questions, please ask now.
- Go on to the next page.

STEP 3. After the first 10 minutes of the test, please advise the students to move to the thinking problems.

- For the thinking problems you have to choose one of the options given and write a letter (A, B or C) in the box provided. There is **only one best answer** for each of the thinking problems.
- Do not guess. If you cannot decide, leave it blank and move to the next question, because you have only a few minutes to complete the questions.
- At the end, if you have time left you are allowed to revise your answers.

- When you finish the test, please remain seated. When the 30 minutes are up, you will be informed.

STEP 4. After 30 minutes, please inform the students that their time is up and request that they put their pencils down. Ask a student to gather the completed tests, while you are ensuring that nobody takes more time to write.

STEP 5. Please complete the post-administration form.

STEP 6. Use the pre-paid envelope to post the surveys back to the School of Education.

Should you have any queries, please do not hesitate to contact me. My e-mail is o.m.ventista@durham.ac.uk. My phone number is 07918519506.

Frequently Asked Questions

What help am I allowed to provide to my students?

During the administration, you can help your students to understand the questions (e.g. you can read the questions to a student with low reading ability or visual impairments) but you should not provide them with hints to respond to the questions. Please do not suggest the correct answer. Even though you can offer help with reading, please do not paraphrase the questions. The vocabulary used has been piloted and confirmed as age appropriate.

What should I do if a student is absent?

The absent students can be given the opportunity to complete the survey form on a different day or at a different time. However, please keep a note on the top of the form that the specific student was absent and please try to ensure similar administration procedures for the student.

Appendix 2c. Post-test administration form which should have been completed by the person administering the assessment (usually the school teacher)

POST-ADMINISTRATION FORM

Please complete this form after the pupils have completed the survey.

If there is more than one Year 5 class in your school, please return one post-administration form for each Year 5 class.

Should you need any assistance to complete this form, please do not hesitate to contact me (o.m.ventista@durham.ac.uk).

School name:	
Number of students in the classroom:	
Number of students who completed the surveys:	
Date of completion	/ /
Any comments or anything to note about the administration	

Does your school implement Philosophy for Children?

YES	
NO	

If yes, for how long has the school been doing Philosophy for Children?

_____

If yes. how often does the specific Year 5 class have Philosophy for Children sessions?

_____

Appendix 2d. Pre-test assessment

Durham University
School of Education

Research project:

PHILOSOPHY FOR CHILDREN

(printed name of the school)

BASIC INFORMATION

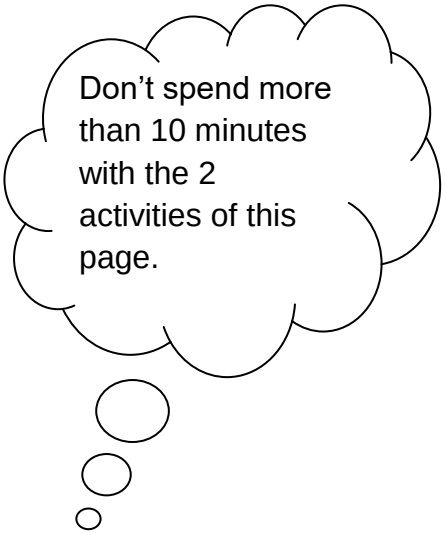
Please complete the boxes.

Sex	
Age (in years)	

INSTRUCTIONS

- Your teacher is going to read some instructions before you start writing. Please listen carefully.
- Please complete all the questions on your own.
- When you have finished answering the questions, please wait quietly until all others have finished.

DO NOT GO ON TO THE NEXT PAGE UNLESS TOLD TO DO SO.



CREATIVITY ACTIVITIES

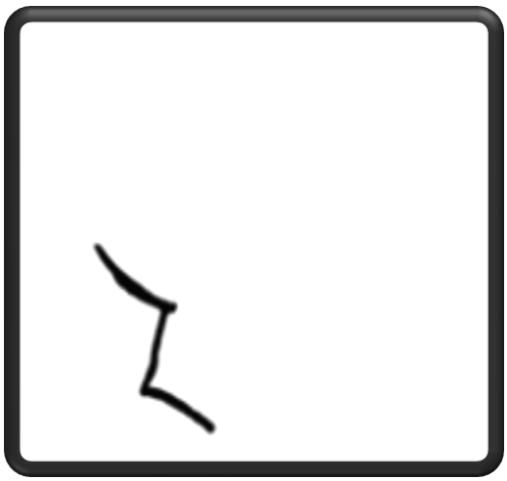
ACTIVITY 1: USES OF OBJECT

Your task is to write down as many different uses as you can for PENCILS. Write down anything that comes to mind no matter how strange it may seem.

_____
_____
_____
_____
_____
_____
_____
_____

ACTIVITY 2: DRAWING

Add lines to the picture below. Try to make it into an interesting object or picture. Try to think of something that your classmates will not think of. Then, think of an interesting title for your drawing.



Write your title in the box below.

Choose one of the options (A, B or C) and write only one of these letters in the box.

THINKING PROBLEMS

PROBLEM 1: DOES JAMES RIDE A BICYCLE?

James says that he rides a bicycle every day. One day you visit him at his house. In the yard, there are some bikes with flat tires.

When you see this,

A. you know that James rides a bike every day.
B. you do not believe that James rides a bike every day.
C. you do not know if James rides a bike every day.

My answer is ☐

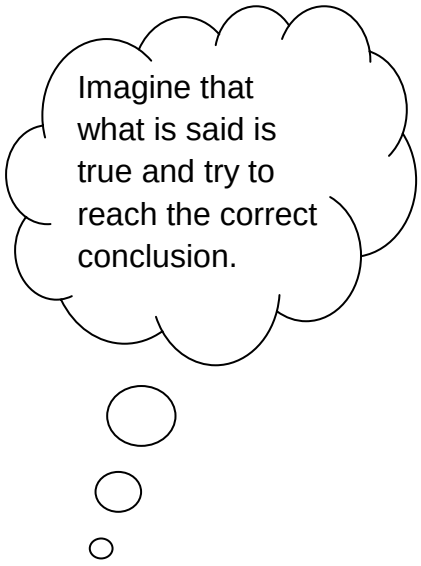
PROBLEM 2: WHO DO YOU BELIEVE?

Nadia is driving in a new city on a Tuesday afternoon. She wants to know about the traffic on the Shaftesbury Avenue. She decides to ask some people to find out. The first person says: 'I drive to my work every day. Whenever I drive on Shaftesbury Avenue in the afternoon hours, I regret it. It is very busy'. The second person says: 'I walk every afternoon from the office to the house. The traffic does not seem to be a problem in this area'.

Whose advice should Nadia follow?

A. The first person
B. The second person
C. Both people

My answer is ☐



PROBLEM 3: THE MEETING

Every time I meet Robert, we go to the cinema to see a film. I did not watch a film yesterday.

This means that

A. I met Robert yesterday.
B. I did not meet Robert yesterday.
C. I might have met Robert yesterday.

My answer is ☐

PROBLEM 4: LISTENING TO CLASSICAL MUSIC

If Kayla's brother is in his room, he always listens to classical music. He plays classical music so loud that Kayla can hear the music in her room. Kayla's brother is in his room. Kayla is not hearing classical music.

This means that

A. Kayla is in her room.
B. Kayla is not at home.
C. Kayla is not in her room.

My answer is ☐

PROBLEM 5: TWO FRIENDS WERE TALKING

'Did you like the orange juice yesterday?', asked Steve. 'I did not drink any juice yesterday, said Charlotte. 'This is not possible. I know that every person drank a glass of orange juice in the party yesterday', said Steve.

This means that

A. there was only orange juice at the party.
B. there was also apple juice at the party.
C. Charlotte did not go the party.

My answer is ☐

PROBLEM 6: THE WEATHER

If it is raining in England in the middle of the night, how likely is to be sunny 24 hours later?

A. It is more likely to be sunny. When it is raining one day, it is more likely to be sunny the next one.
B. There is no possibility of it being sunny.
C. Even if it is raining now, 24 hours later it could be either sunny or rainy.

My answer is ☐

PROBLEM 7: THE THREE BOXES

Tom has three identical boxes. The first box has biscuits, the second box has bars of chocolate and the third has candies. He prepared one label for each box, but he forgotten what it is in each box. What is the least number of boxes he has to open in order to put the correct label on each of the three boxes?

A. One box
B. Two boxes
C. Three boxes

My answer is ☐

END OF THE TEST

Appendix 2e. Post-test assessment

Durham University
School of Education

Research project:

PHILOSOPHY FOR CHILDREN

School: (printed name of the school)

VERSION A

Please complete the boxes.

Sex	
Age (in years)	

INSTRUCTIONS

- Your teacher is going to read some instructions before you start writing. Please listen carefully.
- Please complete all the questions on your own.
- When you have finished answering the questions, please wait quietly until all others have finished.

DO NOT GO ON TO THE NEXT PAGE UNLESS TOLD TO DO SO.

Don't spend more than 10 minutes on the two activities of this page.

CREATIVITY ACTIVITIES

ACTIVITY 1: USES OF OBJECT

Write down as many different uses as you can for BRICKS. Write down anything that comes to mind no matter how strange it may seem.

_____

_____

_____

_____

_____

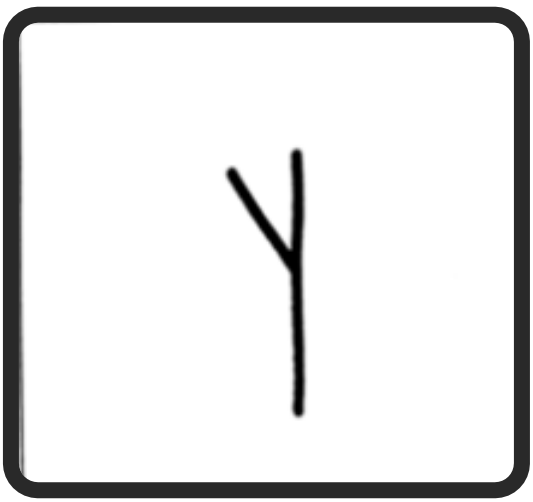
_____

_____

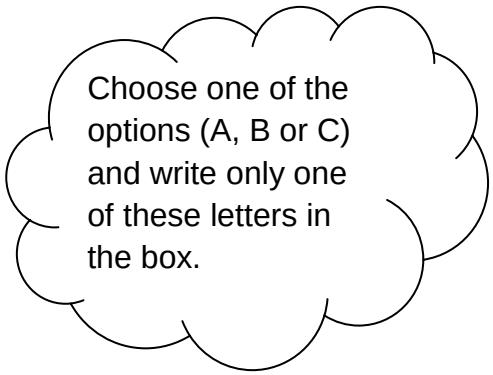
_____

ACTIVITY 2: DRAWING

Add lines to the picture below. Try to make it into an interesting object or picture. Try to think of something that your classmates will not think of. Then, think of an interesting title for your drawing.



Write your title in the box below.



THINKING PROBLEMS

PROBLEM 1: DOES YOUR BROTHER LEARN THE GUITAR?

Today your mother says ‘I think your brother is having secret guitar lessons. I found a ticket from a music concert when I cleaned his room’.

When you hear this, you:

A. think that your brother is having guitar lessons.
B. think that your brother is not having guitar lessons.
C. cannot decide if he is having guitar lessons or not.

My answer is ☐

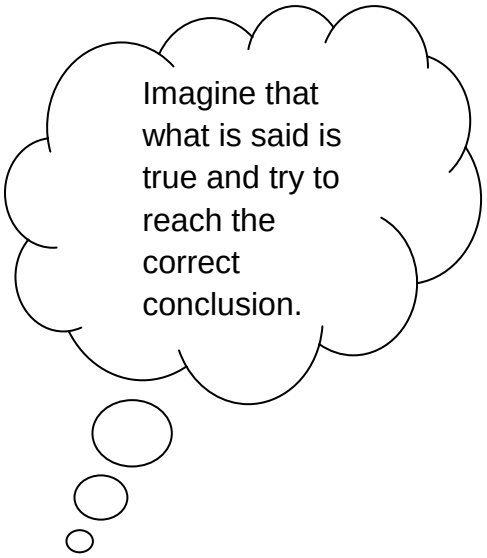
PROBLEM 2: WHO DO YOU BELIEVE?

Sarah has many headaches and she decides to visit a doctor. The doctor asks a series of questions and tells her: ‘You should drink more water’. When she comes out of the doctors, she meets a friend. She explains that she has just been to the doctor for the headaches. Her friend says ‘Every time you are thirsty, you should drink green tea - not water. People say that green tea helps to reduce headaches’.

Whose advice should Sarah follow?

A. The doctor’s
B. Her friend’s
C. Both the doctor’s and her friend’s

My answer is ☐



PROBLEM 3: THE ROAD

If it has rained recently, the road is wet.
The road is not wet.

This means that

A. People do not throw water on the road.
B. It did not rain recently.
C. It might have rained recently.

My answer is ☐

PROBLEM 4: POCKET MONEY

During spring, Peter helps his uncle in return for pocket money. If he saves £500, then he will definitely go to an island in the summer, where he will swim every day.

The summer has arrived. Peter is not swimming today. This means that

A. he saved £500 and decided to spend it differently.
B. he does not swim every day because the sea is cold.
C. he did not save £500.

My answer is ☐

PROBLEM 5: TRAVELLING

Rachel, Oliver and Mark are travelling to different places, each using different types of transport. They go by train, ship and plane. Rachel hates flying. Oliver gets seasick and only has a short distance to travel. What is the most likely transport used by Mark?

A. Train
B. Ship
C. Plane

My answer is ☐

PROBLEM 6: AN ANNOUNCEMENT

Today the Headteacher said: 'Every afternoon there is heavy traffic in front of our school, and a student might be hit by a car. To make sure that no student will be hit by a car, please ask your parents to avoid driving on the road in front of the school entrance in the afternoon' Students reacted differently to this message. Which comment makes more sense?

A. 'The cars are not driven only by our parents. Other people drive on this road, too'.
B. 'The drivers are always careful, so it is unlikely that a student will be hit by a car'.
C. 'The road in front of our school should only be busy in the morning. Not many students are walking in the morning'.

My answer is ☐

PROBLEM 7: GLOVES IN A DARK ROOM

Kayla wants to go out wearing a pair of gloves. In one of her drawers, she has mixed 6 blue and 6 green gloves. It is dark and she cannot see the colours. What is the least number of gloves that she should put into her bag in order to have a pair of gloves of the same colour when she leaves the room?

A. Two gloves
B. Three gloves
C. Seven gloves

My answer is ☐

Did you complete a form like this at the beginning of this school year?

☐ Yes
☐ No
☐ I don't know

END OF THE TEST

Appendix 2f. Comparison between Versions A and B.

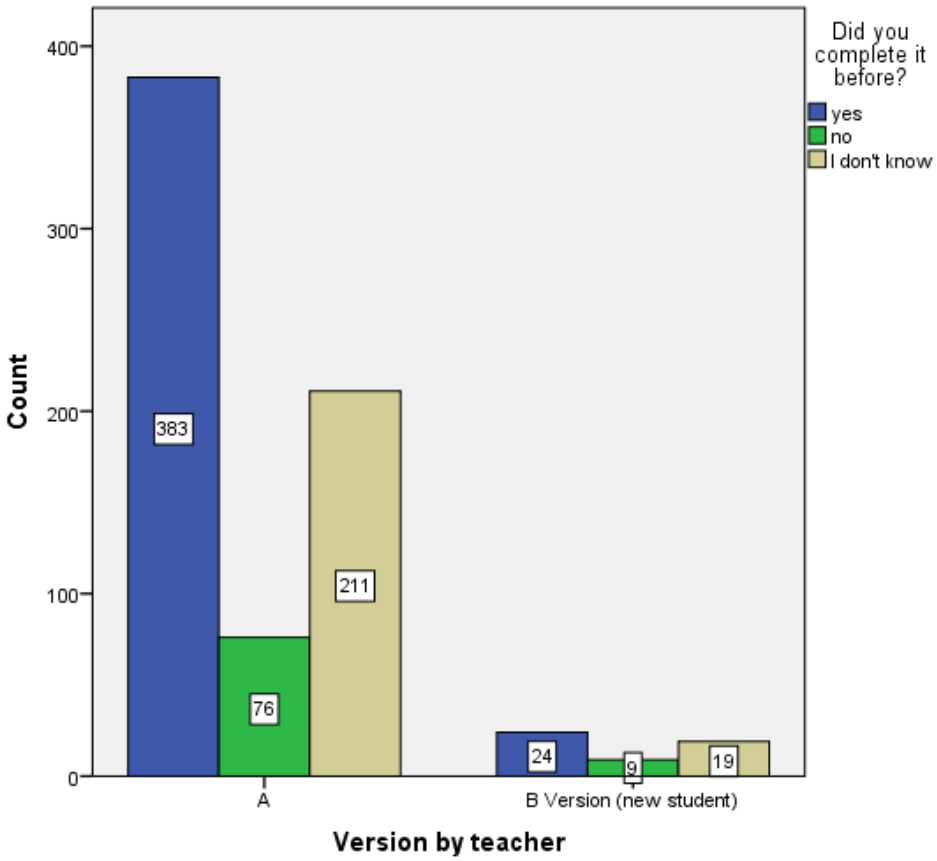
Version A and Version B were basically the same. The only difference was what indicated at the top. I tried to examine the memory of both teachers and students about taking the assessment. The data was anonymous and therefore I could not match their data. The indication of version examines the memory of the teacher as indicated in the post-test administration guide and the final question examines the memory of the pupil. However, there was no reliable indicator found and therefore this indicator was not included in the analysis in order to identify whether the individual had completed the pre-test. Instead the indicator 'intention-to-treat' for the whole school was used.

If a Teacher gave version A, this meant that the teacher believed that the student had completed the survey before and version B that they had not. However, 76 students were given the version A by the teacher, but the students claimed they had not completed this assessment before. Similarly, 24 students were judged as new by the teacher, but they claimed that they completed the assessment before.

Version by teacher * Did you complete it before? Crosstabulation

Count

		Did you complete it before?			Total
		yes	no	I don't know	
Version by teacher	A	383	76	211	670
	B Version (new student)	24	9	19	52
Total		407	85	230	722


Did you complete it before?
yes
no
I don't know
Count
400
300
200
100
0
383
76
211
24
9
19
A
B Version (new student)
Version by teacher

Appendix 3. Chapter 8.

Appendix 3a. Frequency Tables.

Table 1. Frequencies and examples for the Use of Pencil.

Unique Code	Uses for a pencil	Examples of responses	Number of times the category was mentioned
1.	Common Use of Pencil	Writing Scribble Marking Explaining Drawing Colouring Shading Mona Lisa Sketching Taking Notes Crossing out Use a pencil to trace Write with and write on paper or books It's easy to rub out mistakes For a letter to someone Spelling Publish Make a book Books Jotting Doodling Homework Underlining Planning Letters Labelling Projects Maths Computing English Art Science Graph Jobs Working Answer a question	2346

		Tests Poem Poster Patterns Illusions Designing sweet wrappers When taking order at a restaurant as a waiter Voting Recording Smudge Snake Cow Mommy A fossil Window Aliens UFOs Space	
2.	For measurements	Ruler To see if a shelf is straight Measure A piece of non-unit measurement A scaler Comparing Comparing to a pen Counting Tally straws	71
3.	As a rubber	Rubber Erasing (sometimes) Use the end for a rubber	37
4.	As a weapon	Fight Sword Fight Murder Stabbing Stabbing yourself You can make an arrow Crossbow Archery Hitting people Hitting yourself Strike Falling on Breaking things	211

		Kill Killing animals Pretending to be guns Shoot with it Mashing a bug1 A slingshot Catapult Throw at someone Missile Spear Ammunition A lightsaber Attack robbers Beating stick Throwing your pencil to a bird Whacking Stunning Poke eyes out Blowpipe	
5.	As a game	Play with As a toy Games Charlie Charlie Challenge Jenga bricks Dominos Balance on your nose See how much you can balance on top of each one Balance them on your shoulder Balance on your lip Balancing To make them stand on their own Juggling Aiming Aim Having fun Mini pogo stick Dice Give it to your dog	139
6.	For your looks	Brush hair Comb Hair accessories Tie your hair up	131

		To put in your hair Curling hair Parting your hair Making a nose Put it in your ear to look cool Wearing in ear Putting it behind ear when want to look smart For showing off Colouring eyebrows Nail polish Eyeshadow Some people who do strange things stick them in their nose Clowns A brooch Drawing on yourself You could write on your mouth. Lipstick As a moustache Pretend moustache To draw a moustache Moustache maker A man with a beard Use it as a Pinocchio nose	
7.	To itch and scratch	To itch yourself Itching my back Scratching yourself Back scratcher	74
8.	For models	To make a bridge Tower Building a helicopter Car Motor Bike Scooter Road Building Make a tiny see-saw You can use them as a figurine You can make a person You make it into a doll You can dress it up	183

		Clothes Garden gnomes Create Make Make a man You could make a boat Glue them to make a bridge You can use 11,000,000 pencils upwards to be the same height as the twin towers Big Ben You can make stuff with it Bunch of pencils stuck to make a wheel Stick it Glue Clip Use it as a frame for photo Fencing pole Arms for a snowman Feet Nose for snowman Put it on a snowman Bone make illuminati with multiple pencils	
9.	Food-related uses	For a fork Chopsticks Knives Spoon Stirring Stir your tea A mixer Mix Vortex Butter spreader Feed the dog For cooking Pencil ice cream For a candy stick Peeler Also if your ice cream lolly falls you	121

		can put a pencil inside and continue eating it You can make a skewer Kebab	
10.	Put it in the mouth	Biting Put it in their mouths Sucking Eat it To chew Chewing toys Holding chewing gum You can lick it Nibble Taste	130
11.	As a wand	Magic tricks Wand Magic wand Pretending it went through your ear	26
12.	For fire (light and heat)	Setting them on fire for light Fire fuel Make a fire Burn it For a torch Glow in the dark Laser Radiator	37
13.	Be aggressive towards the pencil	Breaking Break if angry Breaking them to make people scared Snap it Snapping Kicking Smash it Bend it Flush them Stick it down your toilet and flush it Put it into the trash Chuck it Grab it Jump on it Shred it	128
14.	To open holes	To rip things Putting holes in	75

		something Pop things Prick something To crack an egg Burst packets of crisps with it Balloon Drilling Bubbles	
16.	For making sound or music	To make noise Drum sticks Music As a metronome To tap on the table To make noise when you drop it	95
17.	Move a pencil	Throw Flinging them To throw and catch Drop Catch Shaking Pushing Pulling Launching Place Twisting it in your hand Fiddling To flip round your fingers If you are bored move it around in your fingers Finger exercise Hand exercise Flicking Flipping Hold it Holding Holding pencils You can hold it Swirling Gripping Twitching Twisting Spinning it on the table Twirling Rolling it	286

		Rolling pin Slide them across the table To carry it around with you Carry it	
19.	As a tool	Unlocking a locked door To pick open locks Key You could open something To open a box Use it as a screw driver For unscrewing screws Fixing As a hammer	34
20.	Teasing	Poke It could be used for annoying your brother Annoy people Waking someone up Jabbing Teasing Pranking Be mean by drawing someone ugly Messing around Being silly Being stupid For stupidness	111
22.	As stress reliever	Stress ball Relieving stress levels by squeezing it or snapping it Stress toy	4
23.	For sports	Javelin throwing See how far you can throw them Karate Football Goalposts Baseball Used for skating Cricket Skiing Tennis	88

		Hockey Physical activities Use for activities Swimming Golf To toughen your hands if you want to do a karate chop Make obstacles Rope Playing darts Darts Snooker	
25.	Reaching and picking something. Moving something with a pencil.	Toothpick Picking nose Picking your nose (don' do that) Take out your ear wax Cleaning your ears Getting dirt out of your finger nails Cleaning out your nails To get things under chairs Dislocate a stone from my bike Reaching To get lead out of a sharpener Pull stuff towards you Taking out tiny things from tight spaces Picking You could pick up stuff Pick up tissue A moving device Moving things Moving something that can stain your clothes Searching	70
26.	Pointing at things and leading the eyes	Pointing at things Show things A pointing stick Show which line you	45

		are reading in a book Reading Following Directing stuff To hypnotise people	
27.	For clocks and compass	You can make a clock with them with a small one and a big one Clock handle Pencil watch Compass	9
31.	Filling	Make piles You can use it to make stacks of pencils To put in pencil cases To fill pencil cases To make your pencil case look full Putting it in your pocket Putting on your car Put it in your shoe To block a hole of a leak Plug a hole Block your ears with Put it in ear Ear socket Plugging the guitar Jamming thing Jam	42
33.	Draw with the back of the pencil (the red end)	Try to draw with the other side of it Colouring in red	7
34.	Engraving	Engraving clay Making details in clay Designing a pumpkin Pumpkin Carving Scraping Chisel	12
35.	Conduct	Using them as conductors Conducting You can use it as the	6

		thing that orchestra holds	
36.	Needlework	Knitting Knitting Needles Using them to make loom bands Sewing	13
38.	Pressing	Press a button Pressing things Pressing buttons on a computer Typing on the computer Tablets Laptops For ipads or phones Internet A way to push tiny buttons Buttons	20
39.	Hold things up	You can balance things on Holding balls Hold things A holder Holding something up Mouth openers Use it as a ring holder Hang things on it Flag Put a note on it and warn someone Lever Elevator Light stand	18
40.	For a distraction	Distracting I can use a pencil as distraction by the throwing it. Thus, the person would get distracted	4
41.	Gardening and digging	Gardening Digging Mining device Pickax Shovel Spade Trees	15

		Plants Planting something out Mud	
42.	Bookmark	Bookmark Use it as a bookmark Keep a page	17
43.	Clean with it	Have it as a toothbrush Toothbrush Cleaner Wash with it using it as a broom stick Tissue Wiping things on it	10
44.	Use your senses	Touch it Feel it For texture Smell Sniff Look at it Staring at You can stare at it	17
45.	Leisure activities (excluding sports)	Acting Clapper Pretend to write Pretend smoking Role playing You can dance with Dance Entertainment Day in the life of a pencil As show Use it to entertain yourself Bring joy to us You can sing with it like a microphone Fake microphone Singing Doing raps Laugh Watching TV at home	27
46.	As a horn	Unicorn Be a unicorn	2
47.	To be helpful	You could send a pencil to a poor	5

		country, so it could help them Helping Share your pencil with other people all of the time Giving it away	
48.	For electricity	Produce electricity with it Electricity For friction Lightning bolt	4
49.	Flying	You can fly on a beach with it Flying Flying with it Stand on the sky Airplane A rocket	11
50.	Handle	Door Handle A grab handle	2
51.	Furniture/ Standing and Sitting on	A table leg Chair You can sit on a pencil Standing on Stepping on	10
52.	Experimenting	Experiment Lab testing Waterproof testing Make salt crystals Crystals	12
53.	Business-related activities	Sell it Trade them for other pencil Buying stuff Shopping Money	6
54.	Drinking	Drink with a pencil Make a straw Bottles	6
55.	To sharpen	Sharpening to make sharpening flowers Sharpening (if you like the noise) Sharpening it until it's tiny You can make shavings with it Making it blunt	13

		Grind	
57.	Floating	Bath toys Floating	5
59.	Thinking	Help you think Thinking Opening door to imagination Remembering Remind Jogging your memory Mind mapping	11
60.	Looking at it and pencil as decoration or artwork	Display could be decoration Hang them for decoration Putting in artwork Decoration Making backgrounds Hang it on something	14
63.	Test your strength	Test your strength Weight	4
66.	Pencil as a living object	Maybe a thing to keep you company Kiss it Pet Hugging	4
67.	As a stick	Walking For a walking stick You can use it if you are a shepherd Used as mini hiking sticks As a stick	6
68.	Cutting	Cutting Cut with it Slice Sawing	8
69.	Tickling	Tickling Use it to tickle your toe	2
71.	Sleep	Sleep Pillow	3
73.	Hiding	For hiding very thin pen knifes in Hide	3
76.	Signal	Signaling	2
81		Recycling	1
82		Attracting Bees	1

83.		As a key ring	1
84.		A door stop	1
85.	Separating	With numbers you can split them up like this 4 |5	1
86.		Fishing rod	2
	Total number of valid answers		4,799
0	Invalid answers		725
	Total number of answers		5,524

Table 2. Frequencies and Examples for the Use of Bricks

Unique Code	Uses for a Brick	Examples of Responses	Frequency
1.	Common Use (Building and construction)	Make a house Doghouse Bedroom Kitchen Wall Stairs A ramp Floor Ceiling Building Schools House(s) Patio Pavement Apartments Roads Tunnels Dead Ends Restaurants Pyramid Shopping Mall Shops Shelter Shed Roof A secret door to a special garden To make a chimney Bridges Build workplaces Build hotels Build hospitals Football stadiums Museums Castles	2728

		House of Parliament Gate Fence Barriers Make a huge tower by stocking them on top of each other Petrol Stations Care homes Churches Work as a builder when you build houses Pillar Making archways Barns Building Birdhouses Big city Village Parks Theme parks Playground Pathway Lighthouse A step Cage Coop Cinema Stage Swimming pool Build treehouses on the ground Play area Playground Fireplace You can use bricks for a treehouse Construction sites Construction Structure Steps Doorstep Fountain Humber Bridge Big Ben Making anything Curving Curve things out of them Shoes Carve them and make heavy shoes with them	

		Food stall Make a table Anvil You could make a chair Chairs Seat Bed Build a bench Furniture Bookshelf Make models Models You can make things out of it Making a totem Making a flag Making Creating Make a stool Make mini houses for squirrels Make a doll house You can make a brick car Airplane Cranes Curve a mini house with it A skateboard for teddy You can make a brick bike A fake tree Super glue Glue stick Stick them to a one and another Sticking Connecting Pegs Star Waterfall Leaves Green house Garden Gardening supplies Manure Grass Make a brick web	
2.	To set limits and protect	Borders Block something off Protecting something	62

		Protection Protecting in a battle Walls for protection Self Defense Defense You use bricks for walls to keep animals in and out To defend yourself Defending your house against robbers Armour They put shelter our people heads It makes people safe To guard something within a wall To separate different areas Any animal enclosure PPE Hat Making a hat Builder's hat Helmet Crown Head	
4.	Using the holes in the brick	House for ants/Bugs Bug Hotel Making a snail house Nest Put pens in the holes Pencil case Pencil pot Pen pot Planting plants in the holes You could make a plant pot Flower pot Holder Pencil holder Egg holder Plant holder Flowerbed Cup holder Marriage rings Sending messages Used to send notes	50
5.	Weapon	Weapon	285

		Weapon if needed Killing spiders For killing people with Killing zombies in a zombie Apocalypse Murder Hit someone over the head with it Hitting little brothers Smash stuff Smash things you don't need Smashing objects Smash glass Breaking Break property Breaking my tooth To throw at glass to break Throwing at a person to knock them out Throwing it at other people Hurt To hurt animals Hurting people A brick sword Catapult Smacking people with them Crushing little objects To crush a bug To use in a sling shot Knock someone out Brick Fights Headbutt	
6.	As a stopper, stabilise and keep things steady	Doorstopper Doors Keeping a door open Keep things still like a door To stop cars from rolling Put them in front of a car's tyre so you can keep it stopped Car stop Anchor Holding things open For holding something in one place	113

		Hold Book holders Keeping veneals steady Stables for horses Stabilise things With brick you can lie your iPad, iPhone , iPod in it so it stands To prop your chin up instead of your hand	
8.	Eating and food-related activities	Eat a brick Trying to eat it if you are dumb Herb-crushing Crack an egg Soften meat Making a saucepan Cook Spoon Utensils Scoop Fake chocolate Plate Trays To eat your food on Put food on	54
10.	Writing (with it or on it)	You could make a book You can write things on a brick Writing on Writing Pencil Pen Chalk Ink Book A red crayon Use another brick to curve some letters Making a poster You could make a whiteboard Notice Board Drawing Drawing on patterns Draw a picture Drawing on the pavement Draw a shape Shapes	129

		Rectangles Draw around it Follow the line	
11.	Play	Playing Playing with A play house Try and play dominos with bricks Children use wooden playing bricks You can play Jenga with them To play rock paper scissors in real life but instead of rock you use a brick Lego Lego Bricks Lego bricks to make Lego people Making things in Minecraft Games Dominos Construct your very own game to use for fun You can make a dice Toys A toy robot head A cuddly toy A game of Tetris Who can carry the most bricks? Bowling balls Who can stack all the bricks first? Using it for a dare on truth or dare Drawing tic tac toe on the floor Juggle them Pretend to be a car on it	135
12.	Art	Make some art with it An art piece Art Arts and crafts Workshop Ferens Art gallery You can also use it for decorating in apartments	214

		Decoration for gates Decoration Painting To use like a stamp on painting I think you can paint it A wall for graffiti Sculpture Statue A thigh of someone Jesus Angel You can imagine a person and try to make it To make a model of a man/woman To build a statue of you Elephant Make a fake cat with it You can make a brick dog Animal Dog Girl Boy Baby Take a picture Colour it Dye Patterns Amy Johnson's moths Display	
13.	Sports	To play tennis with You can create a goal post for football (need two bricks) Karate Chopping for karate Break them with your hand like in karate movies Climbing wall Climbing on them Use it as a cricket bat Swimming Swimming training Ball Frisbee Parkour Ride a brick	57

		A punching bag	
14.	Measurement and Science	Finding the volume of a brick To help find a new formula Break it up to test the materials Make a substitute for milk Experiments Gravity experiment A unit of measurement Comparing Ruler Science lesson on materials See which is heavier Jumping on them and seeing if it breaks	32
15.	Metaphorical Use of the word Brick	Brick a brack is used to say something is not important A person who is not bright An insult Also a villain called bricks A name of a song or a musical Broken heart people Use it in stories	7
16.	Reference to Donald Trump Wall (political use)	Build a wall around Donald Trump Play Donald Trump where you try and make a wall before someone's else Trump’s wall Used to build Donald Trump's wall	7
17.	As weights (including weightlifting)	You can use them to weight lift them, so you get stronger To weight things down Use them a dumbbells Weightlifting Pick Paper weight Weight Paper down in the wind	111

		Holding things down To weigh down a hot air balloon To keep paper where it's supposed to, so it doesn't fly away To make your bags heavy Build up your strength Strong hand To work out Exercise Legs Hand See how much you can hold Tent Stopping a tent from falling Keeping something on the floor Press something light Hot air balloon	
18.	Moving the brick	Throwing Throwing Practice Catching Rock rolling Place it Pulling Fiddle Flip Push Pass Dropping Drop it in a barrel of heated plutonium Drop it off a cliff Carry	81
19.	Make sound	Make music with them by banging them together Music Sound effects Make noises Chants	10
20.	Sit or lay on it	Sit on it Mattress Pillow To sleep on To use as a hard pillow Pillow	16

21.	To relieve stress	Taking your stress out on it Stress toy/throw it against a wall When you are angry smash them To throw at windows when you are angry	6
22.	Fire	Setting fire Make a fire Fire Rub it with a stick to create fire Bonfires You can use it as something that surrounds a fire Surrounding campfires Good to stop fire spreading around you Light a barbeque Boiling Warmth Keep the cold outside A torch Burning materials Lava a barbeque BBQ Making a fake BBQ BBQ holders Oven brick Make an oven Stove	52
24.	Storing	Box (es) I would make a box to collect my favourite things. Secret box Make a box out of bricks A brick box for a birthday Piggy bank Cardboard Cupboard Coffin Bin Bucket Basket Barrel	64

		Tins A file To make presents Seeing what's inside	
25.	As a sharpener	Sharpen Sharpen a pencil	15
26.	Filling	Fill in a room Fill a hole Plugging pipes To stop water flow Sink plug Stack them up Collect them	19
29.	Entertainment	To make a comedy scene Comedy Play script A prop for songs Make a movie about bricks Cosmic bricks in Marvel avengers Toon Cartoon To be silly on	9
30.	Grind	Grind Red dust Make dust Collect the dust Produce more bricks You can break a brick in half Making smaller bricks Crack it then rebuild	16
33.	Scratch	Scratch your back A scratching device Rub your feet Rubbers	11
35.	Drinking	Drink Water bottle Barrel glasses Cup Kettle	24
36.	Cutting	Scissors Chopping Cut Cutting carrots??	9
40.	As a tool	You can use it as a hammer Spatula Spade	27

		Axe Pickax Screwdriver Fixing things Making holes Diggers	
43.	Learning	To learn how to add and subtract You could count them for Maths Numbers Use it for counting Counting Teaching lesson Training course History	17
45.	Tricks/Pranks	Putting them in people's bags for pranks To do magic	4
47.	Use your senses	Look at Observing it Touch Feel You could make a texture Smelling	11
48.	Be aggressive towards the brick	Kick it Slam Punching Crack them Destroy it Poking it Shoot them Bashing them together	13
49.	Money-related uses	Money Payment Sell them Money made from bricks which are worth £1000 Practice handling gold bars	9
50.	Clocks	Clocks Time	7
51.	Brick as a living object	Imaginary Friend Friends To cuddle the night Kiss a brick Make like a pet Family Dress them up	16

		A brick man who helps you with jobs	
52.	Technology	Television Computer Mouse for computer Xbox PlayStation Speakers Laptops iPad Tablets Camera Make a contraption Coffee machine Washing machine Dishwasher Time machine Telephone You could make a Nokia brick Phone cases Robot	38
53.	Environment	Recycle Upcycle Upgrade Smelting Melt to make something new Melt it down to make liquid Polluting World Planet Habitat	20
55.	Become taller and Ladder	Something to stand on Standing on (boost) A decking you can stand on To stand on to make you taller Walk on a brick Podium To reach something you can't reach Knock something of a high shelf Taking down clocks Ladder Stack them up to climb over something	31

		You can help someone if they are trapped Displaying cars	
57.	Balance	You can use them to balance objects on Balance it on your head Balance on To balance Balancing	10
60.	As a stick	Walking stick Sticks Twig	5
62.	To cover	To cover something Secrets Secret stash Hide an egg under if it was Easter or in it Hiding Make a hideout Bury Bury them Lid Blanket A curtain that can move automatically wherever you want the curtain to move won't be cool? Blinds To cover with a blanket in a bed so it looks like you are sleeping	20
64.	Put it in your mouth	Chew Bite Suck Lick it Chomp	6
66.	For looks and hygiene	Makeup brush Eyelashes Nails Clothes Brush hair Washing rock Teeth Brushing teeth	18
67.	For health	Doctors Medicine Cast for arm	3
70.	As an award	Awards Trophy	2
72.	For work	Work	4

		Hard work	
74.	Power and light	Socket Also a convenient brick power generator Light switch Plug socket	4
75.	Sinking	Sink in the sea Sinking something	4
78.	Jumping	Plane bouncers Horse jump To jump over If you want to make an obstacle Create hurdles with them	7
81.	Pointing	Point	2
85.		Flattening paper	1
86.	Punishments	Punishments (drop it on their toe)	1
93.		Alarm trigger	1
100.		Fishing	1
	Total number of valid responses		4598
0	Invalid		418
	Total number of responses		5016

Appendix 3b. Frequency Code Tables.

Table 3. Frequency Code of the categories for the use of pencils.

Frequency Code (number of students who mentioned that category)	Name of the Category	Unique Code
747	Common Use of Pencil	1
196	Move a pencil	17
158	As a weapon	4
136	For models	8
115	As a game	5
107	Teasing	20
104	Put it in the mouth	10
96	For your looks	6
94	Be aggressive towards the pencil	13
84	Food-related uses	9
81	For making sound or music	16

73	To itch and scratch	7
70	For sports	23
69	To open holes	14
66	For measurements	2
60	Reaching and picking something. Moving something with a pencil.	25
42	Filling	31
42	Pointing at things and leading the eyes	26
37	As a rubber	3
31	As a tool	19
31	For fire (light and heat)	12
25	As a wand	11
23	Leisure activities (excluding sports)	45
20	Pressing	38
17	Bookmark	42
17	Hold things up	39

15	Use your senses	44
13	Looking at it and pencil as decoration or artwork	60
13	Gardening and digging	41
13	Needlework	36
12	Engraving	34
11	Thinking	59
11	To sharpen	55
11	Experimenting	52
10	Flying	49
10	Clean with it	43
9	Furniture/ Standing and Sitting on	51
9	For clocks and compass	27
7	Cutting	68
7	Draw with the back of the pencil (the red end)	33
6	As a stick	67
6	Drinking	54

6	Business-related activities	53
6	Conduct	35
5	To be helpful	47
4	Pencil as a living object	66
4	Test your strength	63
4	Floating	57
4	For electricity	48
4	For a distraction	40
4	As stress reliever	22
3	Hiding	73
3	Sleep	71
2	Fishing	86
2	Signal	76
2	Tickling	69
2	Handle	50
2	As a horn	46
1	Separating	85
1	A door stop	84
1	As a key ring	83
1	Attracting Bees	82
1	Recycling	81

Table 4. Frequency Codes of the categories for the Uses of Bricks.

Frequency Code (number of students who mentioned	**Name of the Category**	**Unique Code**

that category)		
690	Common Use (Building and construction)	1
194	Weapon	5
153	Art	12
117	Play	11
103	As a stopper, stabilise and keep things steady	6
93	Writing (with it or on it)	10
93	As weights (including weightlifting)	17
65	Moving the brick	18
58	To set limits and protect	2
53	Storing	24
51	Sports	13
46	Fire	22
44	Using the holes in the brick	4

41	Eating and food-related activities	8
30	Technology	52
27	Become taller and Ladder	55
24	Measurement and Science	14
22	Drinking	35
19	As a tool	40
18	Filling	26
18	To cover	62
17	Learning	43
17	Environment	53
16	Sit or lay on it	20
15	Grind	30
15	Brick as a living object	51
14	As a sharpener	25
11	Be aggressive towards the brick	48
11	For looks and hygiene	66
10	Scratch	33
10	Balance	57
9	Make sound	19

9	Money-related uses	49
8	Entertainment	29
8	Cutting	36
8	Use your senses	47
7	Reference to Donald Trump Wall (political use)	16
7	Jumping	78
6	Metaphorical Use of the word Brick	15
6	To relieve stress	21
6	Clocks	50
4	Tricks/Pranks	45
4	As a stick	60
4	For work	72
4	Power and light	74
4	Sinking	75
3	Put it in your mouth	64
3	For health	67
2	As an award	70
2	Pointing	81

1	Flattening paper	85
1	Punishments	86
1	Alarm trigger	93
1	Fishing	100

Appendix 3c. Prevalence score.

Table 5. Prevalence Score of the categories for the Use of Pencils.

Name of the Category	Frequency Code	Calculation for the category	Prevalence Code
Common Use of Pencil	747	817 – 747	70
Move a pencil	196	817-196	621
As a weapon	158	817-158	659
For models	136	817-136	681
As a game	115	817-115	702
Teasing	107	817-107	710
Put it in the mouth	104	817-104	713
For your looks	96	817-96	721
Be aggressive towards the pencil	94	817-94	723
Food-related uses	84	817-84	733
For making sound or music	81	817-81	736
To itch and scratch	73	817-73	744
For sports	70	817-70	747

To open holes	69	817-69	748
For measurements	66	817-66	751
Reaching and picking something. Moving something with a pencil.	60	817-60	757
Filling	42	817-42	775
Pointing at things and leading the eyes	42	817-42	775
As a rubber	37	817-37	780
As a tool	31	817-31	786
For fire (light and heat)	31	817-31	786
As a wand	25	817-25	792
Leisure activities (excluding sports)	23	817-23	794
Pressing	20	817-20	797
Bookmark	17	817-17	800
Hold things up	17	817-17	800
Use your senses	15	817-15	802

Looking at it and pencil as decoration or artwork	13	817-13	804
Gardening and digging	13	817-13	804
Needlework	13	817-13	804
Engraving	12	817-12	805
Thinking	11	817-11	806
To sharpen	11	817-11	806
Experimenting	11	817-11	806
Flying	10	817-10	807
Clean with it	10	817-10	807
Furniture/ Standing and Sitting on	9	817-9	808
For clocks and compass	9	817-9	808
Cutting	7	817-7	810
Draw with the back of the pencil (the red end)	7	817-7	810
As a stick	6	817-6	811
Drinking	6	817-6	811
Business-related activities	6	817-6	811

Conduct	6	817-6	811
To be helpful	5	817-5	812
Pencil as a living object	4	817-4	813
Test your strength	4	817-4	813
Floating	4	817-4	813
For electricity	4	817-4	813
For a distraction	4	817-4	813
As stress reliever	4	817-4	813
Hiding	3	817-3	814
Sleep	3	817-3	814
Fishing	2	817-2	815
Signal	2	817-2	815
Tickling	2	817-2	815
Handle	2	817-2	815
As a horn	2	817-2	815
Separating	1	817-1	816
A door stop	1	817-1	816
As a key ring	1	817-1	816
Attracting Bees	1	817-1	816
Recycling	1	817-1	816

Table 6. Prevalence Score of the categories for the Uses of Bricks.

Name of the Category	**Frequency Code**	**Prevalence calculation for the category**	**Prevalence Score**

Common Use (Building and construction)	690	738-690	48
Weapon	194	738-194	544
Art	153	738-153	585
Play	117	738-117	621
As a stopper, stabilise and keep things steady	103	738-103	635
Writing (with it or on it)	93	738-93	645
As weights (including weightlifting)	93	738-93	645
Moving the brick	65	738-65	673
To set limits and protect	58	738-58	680
Storing	53	738-53	685
Sports	51	738-51	687
Fire	46	738-46	692
Using the holes in the brick	44	738-44	694
Eating and food-related activities	41	738-41	697
Technology	30	738-30	708
Become taller and Ladder	27	738-27	711

Measurement and Science	24	738-24	714
Drinking	22	738-22	716
As a tool	19	738-19	719
Filling	18	738-18	720
To cover	18	738-18	720
Learning	17	738-17	721
Environment	17	738-17	721
Sit or lay on it	16	738-16	722
Grind	15	738-15	723
Brick as a living object	15	738-15	723
As a sharpener	14	738-14	724
Be aggressive towards the brick	11	738-11	727
For looks and hygiene	11	738-11	727
Scratch	10	738-10	728
Balance	10	738-10	728
Make sound	9	738-9	729
Money-related uses	9	738-9	729
Entertainment	8	738-8	730
Cutting	8	738-8	730
Use your senses	8	738-8	730

Reference to Donald Trump Wall (political use)	7	738-7	731
Jumping	7	738-7	731
Metaphorical Use of the word Brick	6	738-6	732
To relieve stress	6	738-6	732
Clocks	6	738-6	732
Tricks/Pranks	4	738-4	734
As a stick	4	738--4	734
For work	4	738-4	734
Power and light	4	738-4	734
Sinking	4	738-4	734
Put it in your mouth	3	738-3	735
For health	3	738-3	735
As an award	2	738-2	736
Pointing	2	738-2	736
Flattening paper	1	738-1	737
Punishments	1	738-1	737
Alarm trigger	1	738-1	737
Fishing	1	738-1	737

Appendix 3d. Scoring Rubric for Creativity Activity 2.

Scoring for Resistance to Premature Closure

To score this, the marker has to look at the shape the student drew.

The marker gives a score of

0 if the figure is closed in one of the quickest ways with just one line and no further details. This score is also given if the student wrote a letter(s) of the alphabet or number(s). If the student did not add any line in the box, it is marked with 0 (even though the shape remains open).

1 if the student added details inside of the enclosure.

2 the shape is closed but the student added details outside and therefore made it a part of a bigger picture.

3 if there is no closure (the shape is open).

Scoring for the Abstractness of the Title

The marker should mark based only on the text written in the relevant box. If written text is written around the box or in the drawing box, the marker should not consider this text. Particularly, this should be followed when different text is written in the title box and different text in the drawing box. Then, the marker should not choose the one which is scored higher, but the text which is written in the title box.

The rater gives a score of

0 if the title simply names the object depicted. The student simply named the object, or the person depicted it. For example, 'a tie'. There is also the case of naming more than one objects without description (e.g. 'rainbow and unicorn'). There is the case of titles which do not simply state the object, but they include words which do not add information, such as picture or piece (e.g. 'a picture of a nose', 'the piece of chocolate'). The case of naming the object without additional description is included in this case (e.g. 'Max', 'me'). It is important to be noted that the possessive adjectives do not count as additional information (e.g. 'my dad')

1 in three different cases. A) If the title includes some additional descriptive information. This title might include adjectives or gerunds, (e.g. 'happy child', 'wild wolf', 'the writing man'), clarifications (e.g. 'Christmas tree', 'Sunday lunch', 'a rock in a garden', 'crocodile's mouth', 'funny shaped diamond'). B) If the title names a famous person (e.g. 'Queen Elizabeth') or a place (e.g. 'Alps') or/and is seems to be exactly taken by the title of a TV series or show(e.g. 'Pokémon', 'ghostbusters', 'packman', 'Homer Simpson', 'Godzilla'). C) If the title simply names and/or

describes an unrealistic object or a creature which could not be found in reality (e.g. ‘the shape bird’, ‘smiley leaf’, ‘hammer head person’, ‘Mr Egg’, ‘diamond wand’).

2 in this case the title goes beyond what is seen. The title might be humorous and playful (e.g. ‘Guess what it is’, ‘Winner, winner, chicken dinner’, ‘Whaaaat!’ ‘Pretty Girls Oh oh oh’). The student might have invented a word, so this shows increased fantasy. For example a student wrote ‘Map for Lemenia (my made up country)’The title might include an abstract concept (‘happiness’, ‘freedom’, ‘science’, ‘ice cream delight’, ‘nature’s picture’, ‘Justin Bieber’s fairy glore’, ‘Random’, ‘my life of doodle’, ‘idea generator’, ‘the edge of life and death’, ‘in between’) or generalises (e.g. ‘flying’, ‘shocking’, ‘Christmas time’, ‘stormy day’, ‘summer fun’, ‘new beginnings’). The title might also tell a story (e.g. the picture depicts a boy playing football and the title is ‘new hero’, the picture depicts stairs and the title is ‘stairs to nowhere’, the picture shows an alien and it is called ‘an alien invasion’, the picture shows a fish ‘the fish symbol that meets everyone needs to be wealthy’, the picture shows flowers and the title says ‘the amazing flowers that never end’, ‘fruit on the sitting moon’, ‘reaching for the stars’, ‘the light that comes to mind’, ‘the missing star is dead for ever’, the picture shows a tree and it says ‘the tree which never moved’, ‘born to dance’). Generally, for a title to be marked as 2 should not be considered descriptive. For example, it might be a question (e.g. ‘Where is my water?’ ‘No matter what age’

Appendix 3e. Examples of Responses for Activity 2 and their scoring.

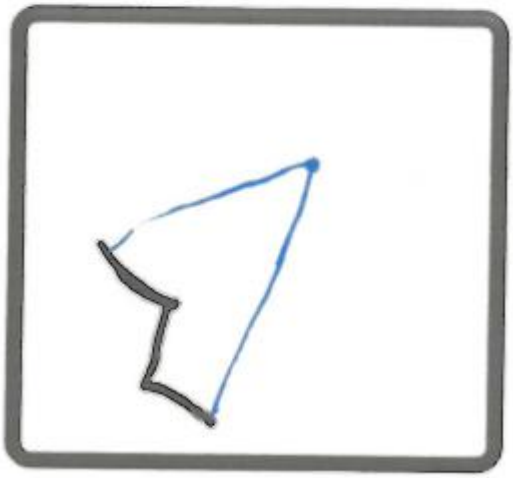


Figure 8.1. Dog's nose.
Questionnaire 108.
The image was scored:
Resistance to Premature Closure (R): 0
Abstractness of Title (A): 0

Figure 8.2. Cool.
Q.401
The image was scored with
R: 0
A: 2

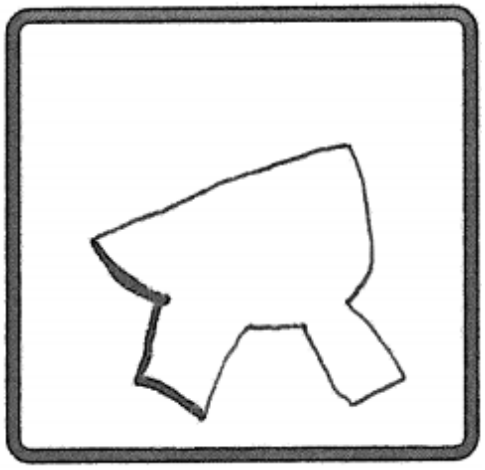


Figure 8.3. Shorts.
Q402.
The image was scored with
R: 0
A: 0

Figure 8.4. Ghostbuster.
Q.101.
The image was scored with
R: 2
A: 1

Figure 8.5. Mr Lightning.
Q.112.
The image was scored with
R: 3
A: 1

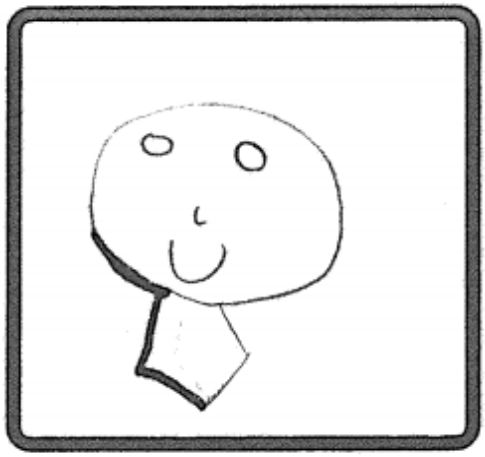


Figure 8.6. The Man with the tie.
Q.405.
The image was scored with
R: 1
A: 0

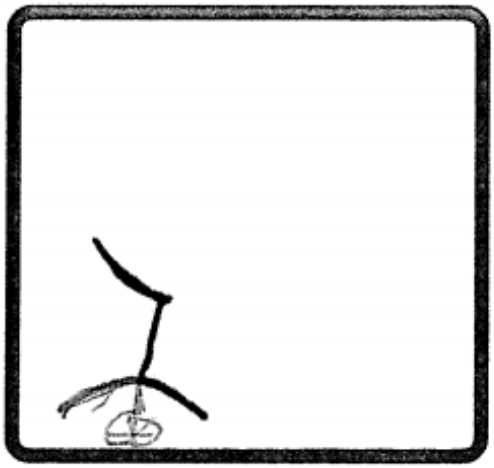
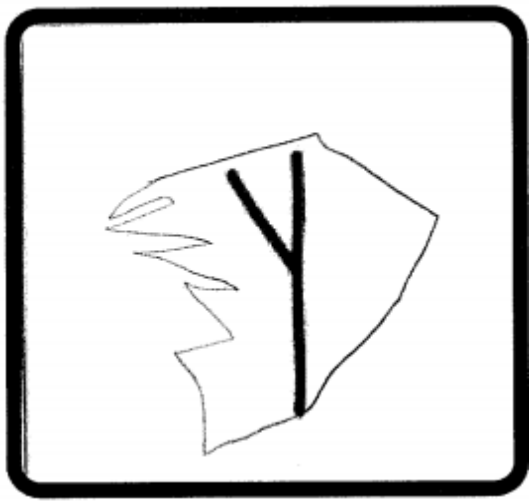


Figure 8.7. Unicycle.
Q.412.
The image was scored with
R: 3
A: 0



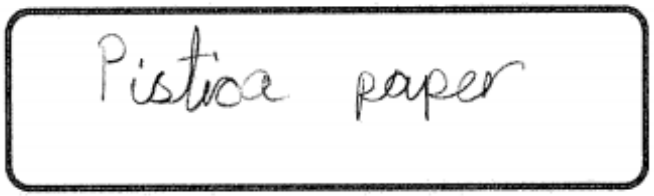


Figure 8.8 and 8.9. Picture and its title.
Q.9003. The student probably means pistachio paper. The image was scored with
R: 0
A: 1

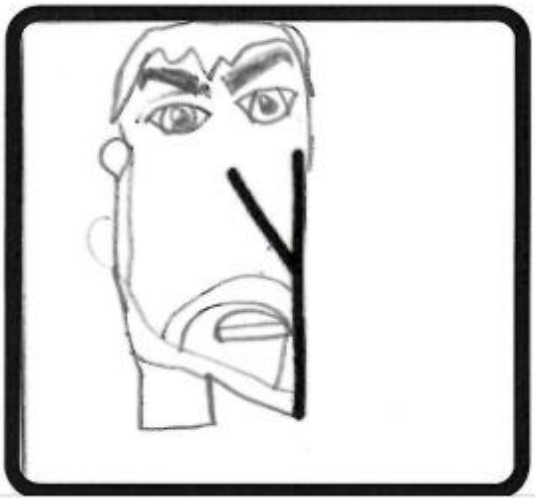


Figure 8.10. The man with no smile.
Q.9075
The image was scored with
R: 2
A: 2

Figure 8.11. Angry monster fun and pigs.
Q. 114.
The image was scored with
R: 2
A: 1

Figure 8.12.
The wind DRAW!
Q.138.
The image was scored with
R: 0
A: 1

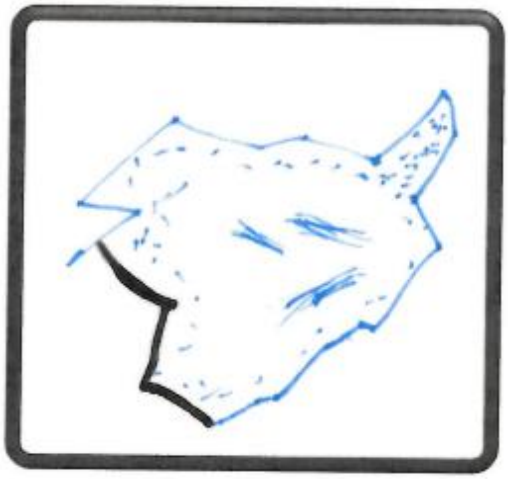


Figure 8.13. Scratches and spots
Q.135.
The image was scored with
R: 1
A: 0

Appendix 4. Chapter 9

Appendix 4a. The two parallel forms used in the piloting.

FORM 1

INSTRUCTIONS

- Please listen to the instructions carefully.
- Do not go on to the next page unless told to do so.
- If you have any questions, please ask now.

REMEMBER! Don't spend more than 10 minutes with the 2 activities of this page.

CREATIVITY ACTIVITIES

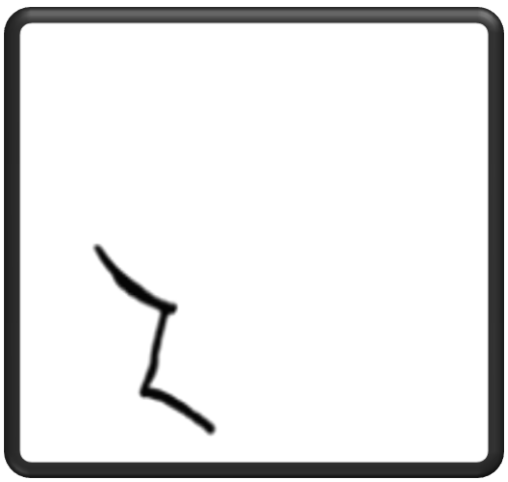
ACTIVITY 1: USES OF OBJECT

Your task is to write down as many different uses as you can for PENCILS. Write down anything that comes to mind no matter how strange it may seem.

_____
_____
_____
_____
_____
_____
_____
_____

ACTIVITY 2: DRAWING

Add lines to the picture below. Try to make it into an interesting object or picture. Try to think of something that your classmates will not think of. Then, think of an interesting title for your drawing.



Write your title in the box below.

REMEMBER! Choose one of the options (A, B or C) and write only one of these letters in the box.

THINKING PROBLEMS

PROBLEM 1: DOES JAMES RIDE A BICYCLE?

James says that he rides a bicycle every day. One day you visit him at his house. In the yard, there are some bikes with flat tires.

When you see this,

A. you know that James rides a bike every day.
B. you do not believe that James rides a bike every day.
C. you do not know if James rides a bike every day.

My answer is ☐

PROBLEM 2: WHO DO YOU BELIEVE?

Nadia is a new driver, and she arrives in a new city. It is afternoon and she wants to know about the traffic on the Shaftesbury Avenue. She decides to ask some people to find out. The first person says: 'I drive to my work every day. Whenever I drive on Shaftesbury Avenue in the afternoon hours, I regret it. It is very busy'. The second person says: 'I walk every afternoon from the office to the house. The traffic does not seem to be a problem in this area'.

Whose advice should Nadia follow?

A. The first person
B. The second person
C. Both people

My answer is ☐

Take for granted that what is said in the box is true and try to reach the correct conclusion.

PROBLEM 3: THE MEETING

Every time I meet Robert, we go to the cinema to see a film.
I did not watch a film yesterday.

This means

A. I met Robert yesterday.
B. I did not meet Robert yesterday.
C. I might have met Robert yesterday.

My answer is ☐

PROBLEM 4: LISTENING TO CLASSICAL MUSIC

If Kayla's brother is in his room, he always listens to classical music. He plays classical music so loud that Kayla can hear the music in her room.

Kayla's brother is in his room. Kayla is not hearing classical music. This means

A. Kayla is in her room.
B. Kayla is not at home.
C. Kayla is not in her room.

My answer is ☐

PROBLEM 5: TWO FRIENDS WERE TALKING

'Did you like the orange juice yesterday?', asked Steve. 'I did not drink any juice yesterday, said Charlotte. 'This is not possible. I know that every person drank a glass of orange juice in the party yesterday', said Steve. This means that

A. there was only orange juice at the party.
B. there was also apple juice at the party.
C. Charlotte did not go the party.

My answer is ☐

PROBLEM 6: THE WEATHER

Tonight in England it is raining at midnight. How likely is to be sunny 24 hours later?

A. There is no possibility of it being sunny.
B. Even if it is raining now, 24 hours later it could be either sunny or rainy.
C. It is more likely to be sunny. When it is raining one day, it is more likely to be sunny the next one.

My answer is ☐

END OF THE TEST

FORM 2

INSTRUCTIONS

- Please listen to the instructions carefully.
- Do not go on to the next page unless told to do so.
- If you have any questions, please ask now.

REMEMBER! Don't spend more than 10 minutes with the 2 activities of this page.

CREATIVITY ACTIVITIES

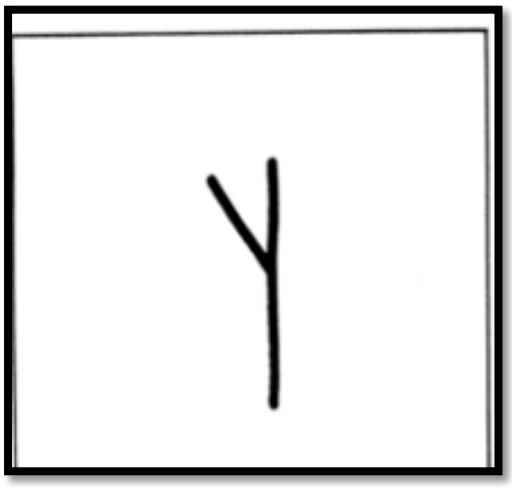
ACTIVITY 1: USES OF OBJECT

Your task is to write down as many different uses as you can for BRICKS. Write down anything that comes to mind no matter how strange it may seem.

_____
_____
_____
_____
_____
_____
_____
_____

ACTIVITY 2: DRAWING

Add lines to the picture below. Try to make it into an interesting object or picture. Try to think of something that your classmates will not think of. Then, think of an interesting title for your drawing.



Write your title in the box below.

REMEMBER! Choose one of the options (A, B or C) and write only one of these letters in the box.

THINKING PROBLEMS

PROBLEM 1: DOES YOUR BROTHER LEARN THE GUITAR?

Your brother wants to learn to play the guitar. Your parents told him to study for school instead and to start learning the guitar next year. Today your mother says 'I think your brother is taking guitar lessons in secret. I found a ticket from a music concert when I cleaned his room'. When you hear this, do you:

A. Agree that your brother has been having guitar lessons.

B. Disagree. You think your brother has not been having guitar lessons.

C. Cannot decide if he takes guitar lessons or not.

My answer is ☐

PROBLEM 2: WHO DO YOU BELIEVE?

Sarah has many headaches. She decided to visit a doctor to deal with the headaches. The doctor asks a series of questions and tells her: 'You should drink more water. ' When she comes out of the surgery, she meets a friend. She explains that she has just come to the doctor for the headaches. Her friend says 'Every time you are thirsty, you should drink green tea - not water. People say that green tea helps to reduce headaches'. Whose advice should Sarah trust?

A. The doctor's
B. Your friend's
C. Both the doctor's and her friend's

My answer is ☐

Take for granted that what is said in the box is true and try to reach the correct conclusion.

PROBLEM 3: THE ROAD

If it has rained recently, the road is wet.

The road is not wet.

Therefore,

A. People do not throw water on the road.
B. It did not rain recently.
C. It might have rained recently.

My answer is ☐

PROBLEM 4: POCKET MONEY

Peter helps his uncle in return for pocket money. Peter saves the money.

If he saves £500, then he will go to an island in the summer, where he will definitely swim every day.

The summer has arrived. Peter is not swimming today. This means that

A. he saved £500 and decided to spend it differently.
B. he does not swim every day because the sea is cold.
C. he did not save £500.

My answer is ☐

PROBLEM 5: AN ANNOUNCEMENT

Today the Headteacher said: “Every afternoon there is heavy traffic in front of our school, and a student might be hit by a car. The school should take care of the students' safety. To make sure that no student will be hit by a car, please ask your parents to avoid driving on the road in front of the school entrance in the afternoon?” Students reacted differently to this message. Which of the three students makes the most sense?

A. “The drivers are always careful, so it is unlikely that a student will be hit by a car”.
B. “The road in front of the school should only be busy in the morning. Not many students are walking in the morning”.
C. “The cars are not driven only by our parents. Other people drive on this road, too”.

My answer is ☐

PROBLEM 6: FOR THE END…LET’S EAT A CAKE!

Rob and Mary love chocolate cakes. They decide to buy a chocolate cake and share it by splitting it into two equal pieces. Both of them should be present when they split the chocolate cake into two pieces, but what should be done to make the sharing as fair as possible? They should both agree in advance that …

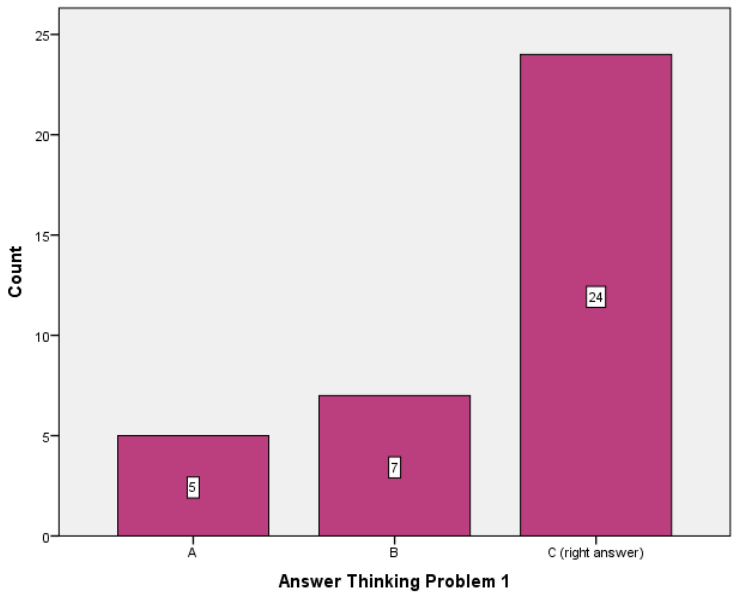
A. Rob will cut the chocolate cake into two equal pieces. Then he will keep one piece and give one piece to Mary.
B. Mary will cut the chocolate cake into two equal pieces. Then she will give one of the two pieces to Rob.
C. Rob will cut the chocolate cake into two equal pieces and then Mary will decide on which piece to take.

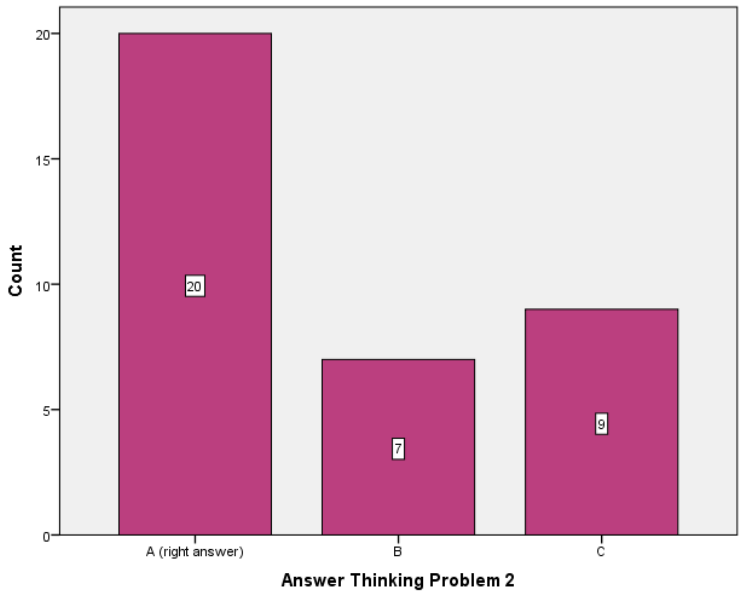
My answer is ☐

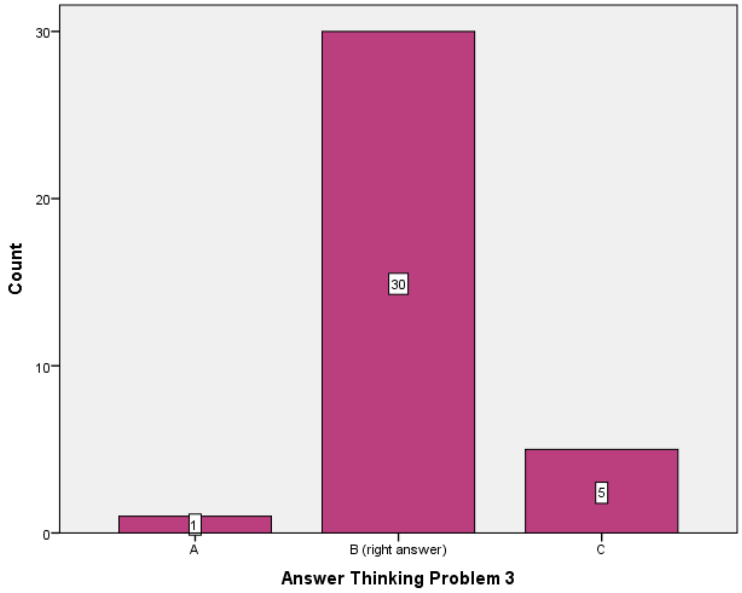
END OF THE TEST

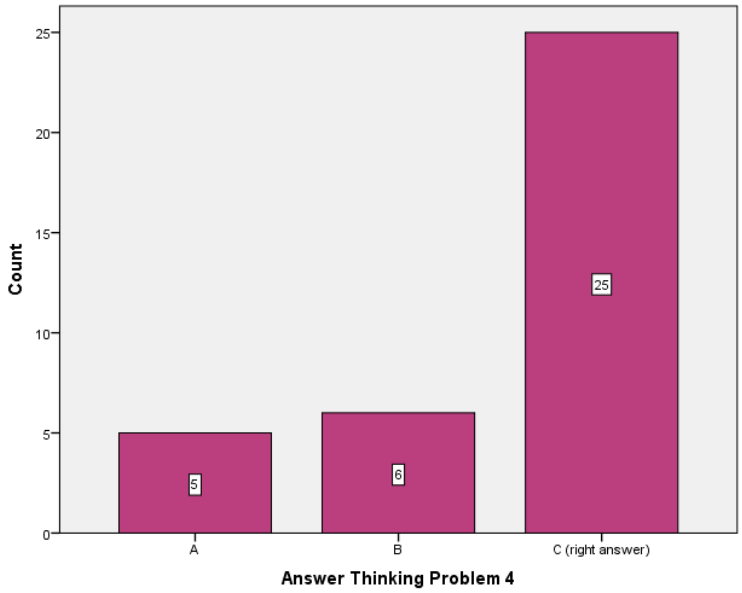
Appendix 4b. Distractors Analysis based on the Pilot Study Data

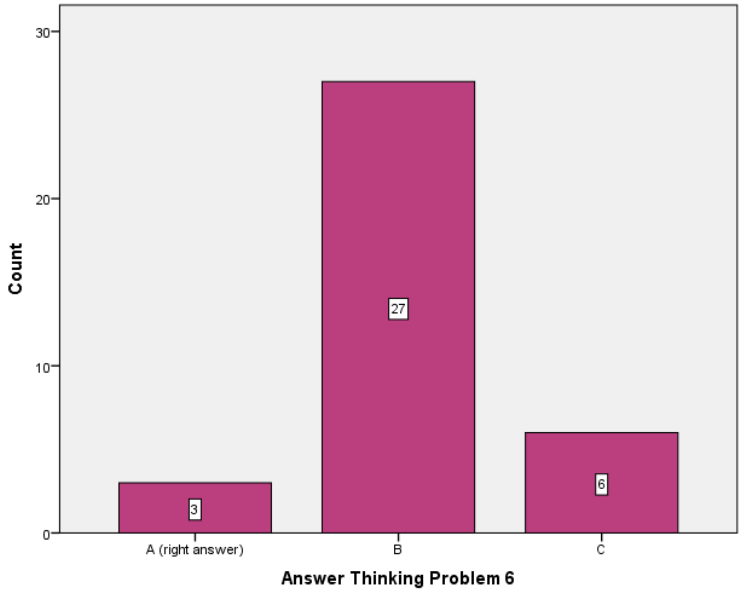
FORM A

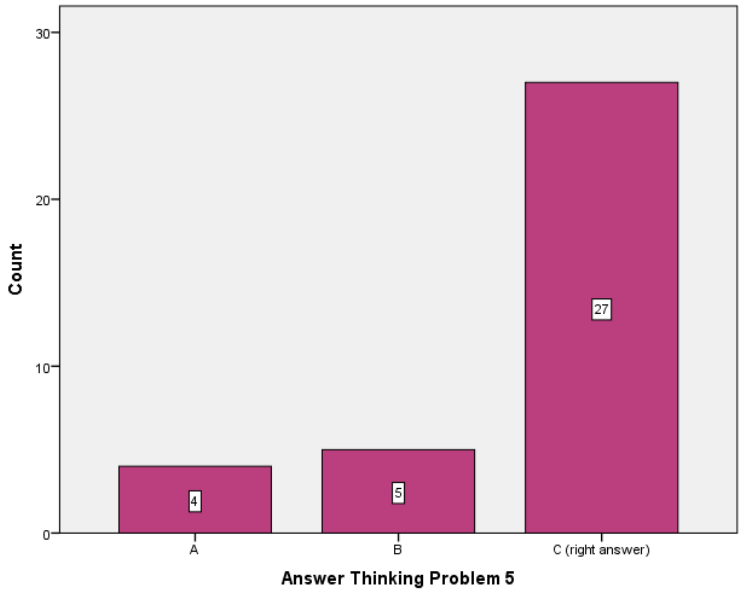




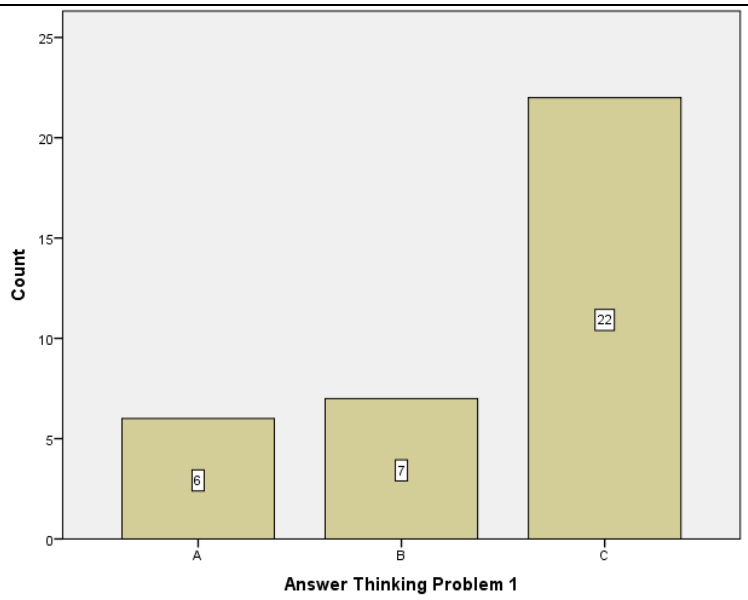

Count
30
20
10
0
30
1
5
A
B (right answer)
C
Answer Thinking Problem 3

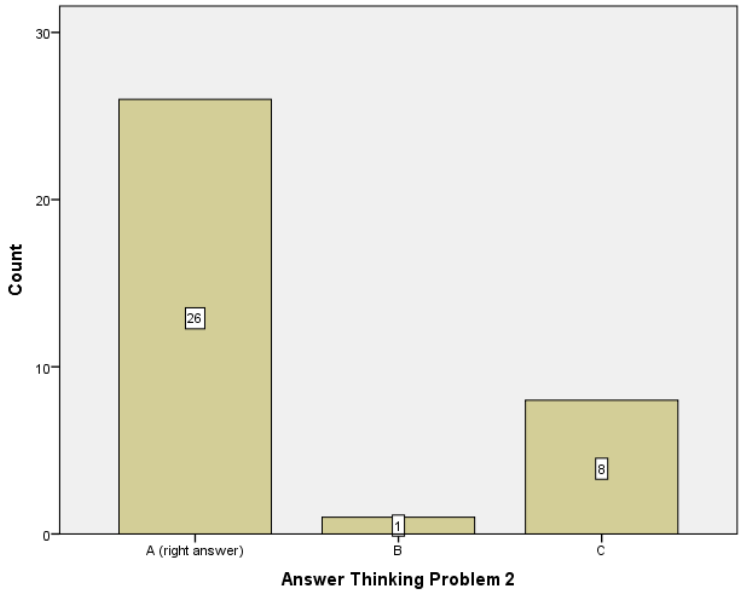

Count
25
20
15
10
5
0
5
6
25
A
B
C (right answer)
Answer Thinking Problem 4

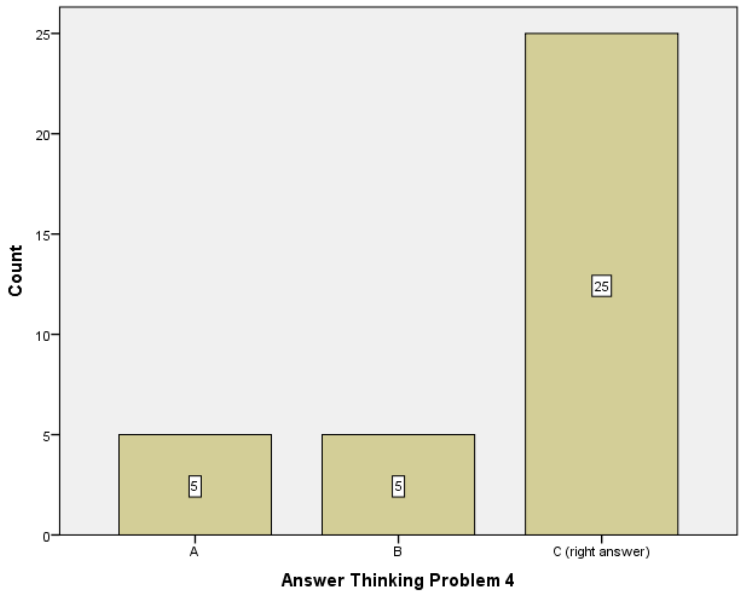

30
20
10
0
Count
27
3
6
A (right answer)
B
C
Answer Thinking Problem 6

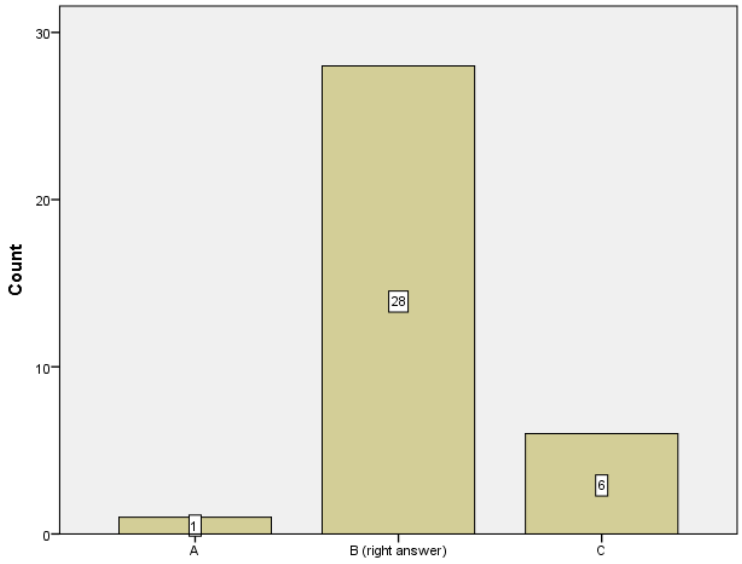

30
20
10
0
Count
27
4
5
A
B
C (right answer)
Answer Thinking Problem 5

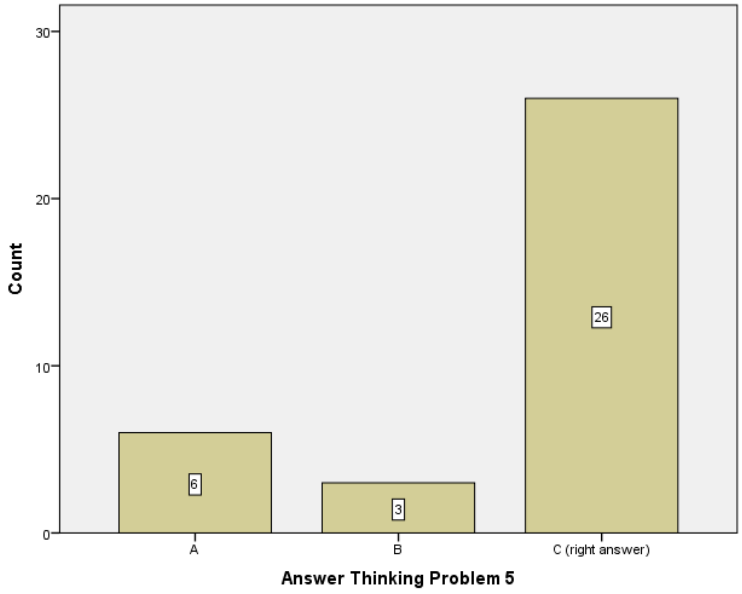
Form B

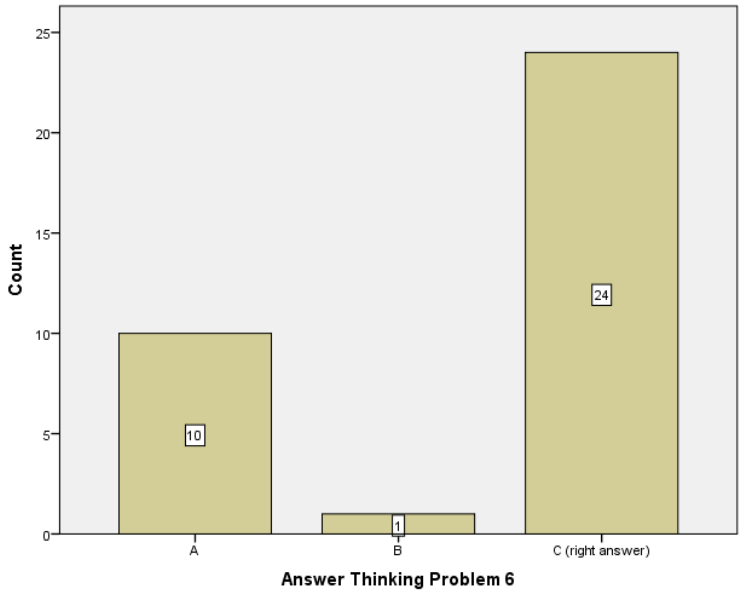

Count
25
20
15
10
5
0
6
7
22
A
B
C
Answer Thinking Problem 1


Count
30
20
10
0
26
1
8
A (right answer)
B
C
Answer Thinking Problem 2


Count
25
20
15
10
5
0
5
5
25
A
B
C (right answer)
Answer Thinking Problem 4


Count
30
20
10
0
28
1
6
A
B (right answer)
C
Answer Thinking Problem 3


Count
30
20
10
0
6
3
26
A
B
C (right answer)
Answer Thinking Problem 5


Count
25
20
15
10
5
0
10
1
24
A
B
C (right answer)
Answer Thinking Problem 6

Appendix 5. Chapter 10.

Appendix 5a. Systematic Literature Review.

Table 7. Research Design, Sample, Results and Effect Sizes of the studies included in the Systematic Literature Review. (presented from the most recently published to the oldest).

Study and quality evaluation	**Research Design (country that research was conducted)**	**Targeted Skills**	**Length of study**	**Follow-up**	**Age**	**Sample Size (N)[3]**	**Attrition[4]**	**Mean (SD) Pre-test**	**Mean (SD) Post-test**	**Effect size[5]**
Rahdar, Pourghaz & Marziyeh (2018) ★★★	Randomly chosen which of the two classes will be in the comparison and in the experimental group (Iran)	Critical Thinking Dispositions (critical openness and reflective skepticism)	12 weeks. (one session per week. Each session lasted 75 minutes)	No	First grade students high-school students	I = 27 C = 27	Sample retained	I = 36.15 (6.37) C= 38 (8.92)	I =46.26 (3.42) C =34.33 (7.83)	1.98
		Self- efficacy (social, emotional and academic)						I =67.22 (9.94) C = 71.85 (15.04)	I = 74.67 (11.49) C = 68.93 (15.75)	0.78

[3] "I" stands for Intervention group and "C" stands for Control group. Similarly, "Final I" stands for the number of students in the intervention group after the attrition (dropouts) and "Final C" stands for the number of students in the control group after the dropouts.

[4] When n.r. is written in this column, it stands for 'no reported'. This means that the attrition was not mentioned in the text retrieved.

[5] Effect size Cohen d (post-tests only). When n.c. is written in this column, it stands for 'no calculated'. This means that the study does not report all of the components to calculate the effect size (sample, standard deviation, means).

Abadi & Akbari (2017) ★★	Study with a comparison group (Iran)	Creativity	10 weeks	No	Nursing students (university students)	I =30 C = 30	Sample retained	I = 134.80 (18.35) C = 123.46 (10.08)	I = 153.23 (12.07) C = 121.36 (7.58)	1.62
Siddiqui, Gorard & See (2017) ★★★★	Study with matched comparison group (United Kingdom)	Communication skills	Intervention lasting from December 2014-June 2016	No	Students Year 4 and 5	Final I= 968 Final C= 1,469	I=131 C=154	I=6.42 (2.81) C=6.03 (2.64)	I=6.25 (2.58) C=6.00 (2.29)	-0.05
		Sociability						I=8.27 (2.62) C=7.97 (2.67)	I=7.79 (2.62) C= 7.67 (2.55)	-0.07
		Cooperation and teamwork						I=7.26 (3.03) C= 6.51 (3.12)	I=7.16 (2.77) C= 6.75 (2.76)	-0.12
		Self-confidence						I=8.15 (2.41) C= 8.16 (2.20)	I= 8.13 (3.90) C=8.00 (2.14)	0.05
		Determination						I= 7.91 (2.81) C= 7.92 (2.66)	I= 7.43 (2.98) C=	0.02

									7.38 (2.63)	
		Social responsibility						I= 7.97 (2.87) C= 7.76 (3.11)	I= 7.67 (2.97) C= 7.77 (2.74)	-0.10
		Well-being						I= 7.45 (3.08) C=7.56 (2.73)	I= 7.22 (2.94) C=7.46 (2.59)	-0.05
		Empathy						I= 7.59 (2.93) C=7.51 (2.77)	I=7.59 (2.63) C= 7.56(2.40)	-0.02
Abbasi & Ajam (2016) ★	Study with comparison group (Iran). Randomisation within the groups	Emotional Intelligence	Twelve 30-minutes sessions	No	Second grade elementary students	I=25 C=25	n.r.	n.r.	n.r.	n.c.
Jahani & Akbari	Study with comparison	Creativity	12 weeks	n.r.	Sixth grade	n.r. (I = 2	n.r.	n.r.	n.r.	n.c.

(2016) (zero ★)	group (Iran)				male students	schools and C =1 school)				
Jahani, Nodehi & Akbari (2016) ★	Study with a comparion group. Randomisation within the groups (Iran)	Moral Judgement	12 weeks	No	Sixth grade girls students because of the stated population in the article (even though in the abstract male students were mentioned)	I= 10 C = 10	n.r.	There is some reporting of mean and SD but it is not clear.	There is some reporting of means and SDs but it is not clear.	n.c.
Säre, Luik, & Tulviste (2016) (zero ★)	Study with comparison group (Estonia)	Verbal Reasoning Skills: Connection between the words (analogy, comparison, contract,	8 months (weekly philosophical discussion)	No	5-6 years old	I = 58 C = 67	n.r.	I = 2.0 (1.9) C= 3.1 (2.3)	I = 5.8 (5.2) C = 2.2 (2.5)	1.47

		justification)								
		Verbal Reasoning Skills: Sense-making explanation-causal connection, understanding about mental stages						I = 0.1 (0.4) C = 1.4 (1.4)	I = 8.3 (4.2) C = 5.0 (3.6)	1.61
		'Because of that' (justification)						I = 0.6 (1.2) C = 2.8 (2.4)	I =7.6 (3.8) C=4.9 (3.6)	1.66
		Talkativeness						I = 80.90 (32.7) C= 90.2 (32.8)	I = 212.1 (156.9) C= 133.7 (90.8)	0.95
Shatalebi & Hedayati (2016) ★★★	Study with random allocation of participants in groups (Iran)	Psychosomatic disorders	12 weeks (12 P4C sessions. Each of the sessions lasted for an hour)	No	9-11 years old (in the abstract and title it mentions only 'boys', in the section	I = 23 C= 22	Sample retained	I = 16.9 (3.65) C= 7.08 (1.63)	I = 3.56 (1.55) C = 6.48 (1.75)	-5.47

					about the sample it mentions only 'female' students)					
Tian & Liao (2016) ★★	Study with a comparison group	English learning anxiety	10 weeks (100 minutes per week)	No	Students aged 16-17 years old (engineering major)	I = 29 C =33	Sample retained	I = 42.41 (n.r) C= 43.64 (n.r)	I = 40.83 (n.r.) C = 39.00 (n.r.)	0.36[6]
		English learning motivation						I= 63.79 (n.r.) C = 98.21 (n.r.)	I = 51.62 (n.r.) C = 51.27 (n.r.)	2.58[7]
		Reading comprehension						I = 7.03 (n.r.) C = 4.85 (n.r.)	I = 12.28 (n.r.) C = 10.67 (n.r.)	-0.19[8]
Youssef,	Study with	Reading	Six	6	Year 6	Final	I=13	I= 41.96	I=47.9	0.32

[6] The study reports only paired SD (I =7.84 and C = 9.19). If these are used for both pre-test and post-test, this is the effect size that occurs. However, this result should not be considered directly comparable with the others in the table because there was a compromise in its calculation.

[7] The study reports only paired SD (I= 10.94 and C = 15.37). If these are used for both pre-test and post-test for the groups, this is the effect size that occurs. However, this result should not be considered directly comparable with the others in the table because there was a compromise in its calculation.

[8] The study reports only paired SD (I = 3.16 and C = 3). If these are used for both pre-test and post-test for the groups, this is the effect size that occurs. However, this result should not be considered directly comparable with the others in the table because there was a compromise in its calculation.

Campbell & Tangen (2016) ★★★	comparison group (Australia)	Comprehension	months (June 2011-December 2011)	months (June 2012)	students (10-12 years old)	I=117 Final C=105	C=5	(11.67) C=47.68 (12.30)	2 (12.38) C=49.60 (13.49)	
		Interest in Maths				Final I=118 Final C=105	I=12 C=5	I=28.16 (7.81) C=27.07 (8.35)	I=26.57 (8.39) C= 28.49 (8.44)	-0.37
		Self-esteem				Final I=118 Final C=105	I=12 C=5	I= 28.93 (4.24) C=28.98 (4.20)	I=28.77 (4.48) C= 30 (4.09)	-0.28
		Pro-social behaviour				Final I=116 Final C=104	I=14 C=4	I=7.77 (1.76) C=7.53 (1.83)	I=8.06 (1.66) C= 7.74 (1.67)	0.46
		Emotional well-being				Final I=115 Final C=104	I=15 C=4	I= 3.53 (2.06) C= 3.06 (2.24)	I= 3.06 (2.44) C= 2.90 (2.22)	-0.14
Abaspour, Nowrosi & Latifi	Study with comparison group	Awareness	15 sessions	No	Female students (12-14	I = 15 C = 15	Sample retaine	I = 69.93 (10.51)	I=78.26 (9.93)	0.62

(2015) ★★	(Iran).				years old)		d	C= 66.80 (10)	C= 68.53 (12.54)	
		Realistic acceptance						I=20.06 (9.48) C=19.80 (11.79)	I=21 (8.8) C= 25.01 (10.10)	-0.42
		Disappointment						I= 16.3 (4.82) C= 15.8 (6.8)	I = 13.33 (6.46) C = 19.93 (6.39)	-1.15
		Grandiosity						I=20.20 (7.00) C= 23.53 (5.71)	I = 17.53 (4.4) C = 23.86 (5.12)	-0.53
		Instability						I= 23.73 (3.78) C= 29.2 (5.26)	I = 23.26 (4.7) C = 30.8 (5.03)	-0.44
		Impression management						I= 18.8 (3.58) C= 19.8 (3.48)	I = 14.8 (7.36) C =	-0.41

									17.8 (3.93)	
Cooke (2015) ★★★	Study with a control group (United States). Randomisation within the groups	Critical thinking	6-8 P4C sessions in total	No	6th grade students	I=15 C=9	I=1 C=1	I= 9.53 (n.r.)[9] C = 8.00 (n.r.)	I= 14.67 (n.r.) C = 7.56 (n.r.)	n.c.
		Construction						I = 1.73 (0.46) C= 1.78 (0.44)	I = 3.0 (0.76) C =1.67 (0.50)	2.40
		Cogency						I = 1.87 (0.55) C =1.44 (0.53)	I = 2.80 (0.86) C = 1.56 (0.53)	1.33
		Adaptability						I = 2.07 (0.96) C= 1.33 (0.71)	I =3.27 (1.00) C=1.11 (0.33)	1.83

[9] The study reports the results of each question individually and reports SD for these 5 items (construction, cogency, adaptability, metacognition- 2 items). However, this does not seem particularly useful. First of all, even though the study discusses critical thinking the items does not make clear how they match the aspects of critical thinking. Moreover, there are two items to measure meta-cognition and they appear to be reported separately.

		Metacognition (Q4): Do you ever think about how you think or how you reason?						I = 1.87 (0.63) C = 1.56 (0.53)	I = 2.73 (0.59) C= 1.56 (0.53)	1.44
		Metacognition (Q5): Do you think a person could learn to 'think better'?						I = 2.00 (0.65) C = 1.89 (0.78)	I = 2.87 (0.74) C = 1.67 (0.50)	1.60
Fair et al. (2015a) ★★★	Study with a comparison group with randomisation of teachers within the same school (Texas)	Reasoning skills	22-26 weeks for 7th graders and 4-10 weeks for 8th graders	No[10]	7th and 8th graders (12 and 13 years old)	I = 363 C =177	n.r.	I = 102.19 (32.69) C= 93.86 (36.99)	I= 119.38 (31.74) C= 104.23 (35.32)	0.20
Fair et al. (2015b)	Study with a comparison	Reasoning skills	3 years follow up study	Yes	Only the initial 7th graders	Final I = 133 Final C	I =53 C = 29	I = 100.09 (30.41)	I = 122.53 (35.25	0.34

[10] The study does not mention follow up. However, the same authors report the post-test after three years folllow-up (Fair et al., 2015b).

★★	group (Texas)				(but now 15-16 years)	= 50		C= 89.60 (37.40)	) C= 100.26 (39.09)	
Gorard, Siddiqui & See (2015) ★★★★★	Randomised Control Trial. Randomisation at a school-level. (England)	Reading	1 year (December 2012-January 2014)	No	Year 5 pupils	I=772 C=757	Sample retained	I= -0.08 (1.01) C=0.08 (0.98)	I= -0.02 (1.01) C=0.02 (0.99)	0.12
		Maths						I= -0.09 (1.04) C=0.08 (0.95)	I=-0.04 (1.01) C=0.04 (0.99)	0.09
		Writing						I=-0.07 (1.03) C=0.07 (0.96)	I=-0.05 (1) C=0.06 (1)	0.06
		Reasoning skills				Final I=1,366 Final C=1,455	Drop-out I=184 C=154	I= 94.37 (11.24) C=95.20 (11.19)	I=96.59 (12.26) C=96.90 (11.90)	0.04
Tok & Mazi	Study with a	Reading Comprehensio	One academic	No	5th graders (10 to 11	I=37 C=37	Sample	I= 25.89 (6.13)	I= 28.08	-0.09

(2015) ★★	comparison group (Turkey)	n	year		years old)		retained	C= 25.16 (6.01)	(5.21) C= 27.83 (4.90)	
		Listening Comprehension						I= 22.02 (6.17) C=23.48 (6.50)	I=24.24 (6.38) C=23.72 (7.38)	0.30
Colom, Moriyon, Magro & Morilla (2014) ★	Study with a comparison group (Spain)	Cognitive ability and Personality	4 years	Planned Longitudinal design. Until the students were 16 years old	Data obtained at two different points. When students were 2nd and then 6th grade (8 to 12 years)	I =281 C=146	n.r.	n.r.	n.r.	n.c.
Nia (2014a) ★★	Study with a comparison group (Iran)	Anger (Overall wrath index)	20 sessions	No	First grade of high school	I=30 C=30	Sample retained	I = 41.6 (41.6) C = 39.7 (39.7)	I =33.3 (33.3) C = 48.1 (48.1)	-0.41
Nia (2014b) ★★	Study with a comparison group	Public Self-esteem	20 sessions	No	First grade of high school	I=30 C=30	Sample retained	I = 47.19 (3.06) C =	I = 2.18 (3.69) C =	-12.9

	(Iran)							20.4 (4.44)	22.03 (3.02)	
		Social Self-esteem						I =6.03 (1.88) C =5.8 (1.42)	I = 2.15 (6.3) C =5.1 (1.65)	-0.92
		Family Self-esteem						I = 6.4 (1.45) C =6.3 (1.39)	I =6.77 (1.74) C =5.37 (1.65)	0.83
		Educational Self-esteem						I =5.17 (1.58) C =5.6 (1.22)	I = 5.8 (1.04) C = 5.03 (1.54)	0.88
Pourtaghi, Hosseini & Hejazi (2014) ★★	Study with a comparison group (Iran)	Creativity: fluency	5 sessions for 75 minutes	No	Boys only (second grade of secondary school)	I =16 C =16	Sample retained	I = 15.44 (3.75) C = 18.56 (4.66)	I = 38.06 (23.77) C = 23.56 (10.75)	1.32
		Creativity: flexibility						I = 12.94 (3.30) C = 16 (4.29)	I=21.13 (7.85) C=17.19 (7.46)	1.16

		Creativity: innovation						I = 20.81(6.23) C = 26.19 (9.06)	I=47.5 (27.82) C=23.94 (22.39)	1.55
		Creativity: elaboration						I = 63.56(22.1) C =69.5 (17.46)	I=148.38 (51.33) C=91.31 (40.41)	1.77
Giménez-Dasí, Quintanilla & Daniel (2013) (zero ★)	Study with a comparison group (Spain)	Emotion Comprehension	30 sessions (one academic year from October since May)	No	4 years old	I =18 C=9	n.r.	I = 4.42 (1.57) C = 4.38 (1.30)	I =4.94 (1.55) C=6.22 (1.99)	-0.83
					5 years old	I=14 C=19		I=5.57 (1.16) C=5.56(1.20)	I= 7.43 (1.75) C= 6.22 (1.17)	0.91
		Knowledge about Strategies for Interaction with			4 years old	I=18 C=9		I=2.53(1.07) C=2.94 (1.75)	I= 3.49 (1.64) C=3.17	0.49

		Classmates							(1.62)	
					5 years old	I=14 C=19		I=2.73 (1.13) C=2.83 (1.1)	I=5.91 (1.31) C=3.39 (1.11)	2.27
Lam (2012) ★★★	Study with comparison group. Randomisation within the groups (Hong Kong)	Reasoning skills	Twice a week 90 minutes sessions for 16 weeks	No	Secondary I	I=14 C_1=14 C_2=14	Sample retained	I= 27.14 (5.56) C= 30.50 (6.10)	I= 34.71 (5.68) C= 34.57 (5.23)	0.62
Reznitskaya et al. (2012) ★★★ ★★	Study with comparison group. Randomisation within the groups (New Jersey)	Transfer of argumentation development (different variables. Only the student related are reported here) Elaborated Reasoning	Once per week for 12 weeks. 40-minutes session.	No	5th grade	Final I = 135 Final C = 125	I = 3 C = 0	Pre-test Reading Comprehension I = 38.9 (7.8) C = 39.2 (6.5) Persuasive Essay I = 2.9 (1.1) C = 2.8 (1.0)	I = 22.2 (5.4) C = 11.0 (9.3)	1.49
		Student Questioning							I = 5.3 (3.3) C = 2.7 (3.6)	0.76
		Elaborated							I = 0.1	-2.62

		Description. Recall (students responding to the question 'what happened?'							(0.5) C = 9.9 (5.4)	
Hedayati & Ghaedi (2009) ★★★	Study with comparison group. Randomisation within the groups (Iran)	Interpersonal relationships	Twelve 90-minutes sessions	4 months follow-up	3rd to 5th graders	I = 88 C= 102	Sample retained	n.r.	Not clear reporting	n.c.
Marashi (2008) ★	Study with comparison group. Randomisation within the groups (Iran)	Reasoning skills	Eleven 70-minutes sessions	No	8th grade students. Only boys	I=30 C=30	n.r.	I=31.40 (4.34) C=30.76 (5.17)	I=35.36 (3.93) C= 29.83 (5.43)	1.02
Topping & Trickey (2007) ★★	Study with a comparison group. Randomisation within the groups with two schools participating (United	Cognitive gains (overall)	16 months (one hour per week)	2 years	10 years old	I=105 C=72	n.r 'not significant attrition' (p.277)	I=99 (13.1) C= 101.3 (12)	I= 105 (14.1) C= 99.4 (13.2)	0.60

	Kingdom)									
Trickey & Topping (2006) ★	Study with a comparison group. Randomisation within the groups. (United Kingdom)	Self-esteem[11]	7 months (one hour per week)	No	11-12 years old	I=119 C=52	n.r.	I=71.37 (13.50) C=70.36 (14.2)	I=72.6 (12.5) C=72. 88 (10.7)	-0.10
Acedo Lizarraga et al. (2003) ★★★	Study with comparison group. Randomisation within groups (Spain)	General Intelligence	Portfolio intervention (Philosophy for Children is a part of this intervention) 120 hours	2 years	13 years old	I=20 C=20	Sample retained	I=105.1 (10.92) C= 105 (5.48)	n.r.	n.c.
		Cognitive Flexibility						I=40.75 (15.06) C=34.35 (14.15)	n.r.	n.c.
		Metacognitive Strategies						I= 39.70 (15.56) C=34.45 (17.20)	n.r.	n.c.
		Academic Achievement						I= 5.35 (1.22) C=5.45 (0.75)	n.r.	n.c.
Schleifer et al. (2003) ★★	Study with comparison group (Montreal	Moral Autonomy	Weekly intervention for about an	No	Kindergarten students (5-year-	I =39 C =42	Sample retained	I = 3.53 (n.r.) C= 3.53 (n.r.)	I = 4.03 (n.r.) C=	n.c.

[11] This study also involved a social skills assessment for the pupils completed by the teachers.

	area, Canada)		hour (from October until April)		old)				3.77 (n.r.)	
		Judgment						I = 4.47 (n.r.) C= 7.53 (n.r.)	I = 9.10 (n.r.) C= 9.28 (n.r.)	n.c.
		Empathy						I = 1.26 (n.r.) C= 1.23 (n.r.)	I = 1.77 (n.r.) C= 1.19 (n.r.)	n.c.
		Emotion-Recognition						I = 17.39 (n.r.) C= 17.00 (n.r.)	I = 19.33 (n.r.) C= 19.02 (n.r.)	n.c.
Jo (2001) ★★	Study with comparison group (Korea)	Meaning Construction	24-week programme (April-July and September-November)	No	Kindergarten students(5-year-old)	I=27 C=27	Sample retained	I= 2.44 (1.81) C= 2.74 (1.52)	I=3.70 (1.28) C=2.85 (1.16)	0.79
Sprod (1998) ★★	Study with comparison group (United	Science reasoning tasks	An academic year (weekly	No	Year 7 students (11-12 years old)	I = 25 C=29	Sample retained	I=5.28 (1.08) C=5.50 (1.22)	I=6.57 (0.82) C=6.29	0.51

	Kingdom)		70-minutes sessions)						(0.71)	
Schleifer & Poirier (1996) (zero ★)	Study with a comparison group (Canada)	Stereotypic Attitudes and Respect for others	An academic year (once per week)	No	Second year classes	N =26	n.r.	n.r.	n.r.	n.c.
Fields (1995) (zero ★)	Study with comparison group (United Kingdom)	Academic achievement, reasoning skills, self-image, behaviour, motivation	2 years	No	7-8 years old	N=123	n.r	n.r.	n.r.	n.c
Sasseville (1994) (zero ★)	Study with a comparison group (Canada)	Self-esteem and logical skills	5 months	No	3rd to 6th graders	I=124 C=96	n.r.	n.r.	n.r.	n.c.
Williams (1993) ★★	Study with a comparison group (United Kingdom)	Reading ability	One academic year October - June (27 one-hour sessions)	No	Year 7 pupils (11-12 years old)	I = 15 C = 17	Sample retained	I = 91.5 (n.r) C = 89.3 (n.r.)	I = 94.5 (n.r) C = 89.4 (n.r)	n.c.
		Intellectual confidence				I =14 C=14		I = 41.9 (n.r) C=44.1 (n.r)	I = 47.2 (n.r) C =	

									44.7 (n.r.)	
Slade (1989) ★★	Study with a control group (Australia)	Reasoning skills	Twelve 2-hour sessions	No	Grade 7 and female students only. Top Year 7 Math Group	I = 15 C = 15	Sample retained	I = 38.95 (8.50) C = 39.34 (7.05)	I = 45.29 (3.81) C = 42.33 (3.77)	0.48
					Lowest Year 7 Math Group	I = 10 C =10		I = 34.10 (5.42) C = 30.57 (4.83)	I = 39.20 (7.08) C = 34.29 (4.89)	0.25
Russell (1988) ★	Study with a matched comparison group (United States)	Verbal reasoning related to defining art	40 minutes instruction (twice per week)	No	5th and 6th grade students	I=26 C=25	n.r.	n.r.	n.r.	n.c.
Banks (1987) ★	Study with a control group (United	Reading, Language Arts, Maths[12]	One academic year	No	Primary School pupils (Grades	I = 139 C= 133	Teacher attrition	n.r.	n.r.[14]	n.c.

[12] The programme reports only the overall gain scores for the overall California Achievement Test which involves these three areas.
[14] The pre-test and post-test means are not reported. However, gain scores are reported.

	States)				2-5)		reported[13]			
Jenkins (1986) ★★	Study with a comparison group (United Kingdom)	Reasoning skills	One academic year	No	12 years old pupils	I=30 C=30	Sample retained	I= 25.69 (10.59) C=29.66 (9.27)	I =36 (8.61) C= 34.03 (8.69)	0.64

[13] The study does not report the student attrition. It reports only the teacher attrition. This does not enable the reader to know the number of students who dropped out, because the author reports pre-test results only from the students who also got the post-test (N =272).

Bibliography

Abadi, F. H. D., & Akbari, A. (2017). Studying the Effect of Community of Inquiry in Philosophy for Children Program on Creativity Improvement among Nursing Students. *Int J Sci Stud*, 5(4), 494-500.

Abaspour, N., Nowrosi, 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), 282-286.

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A. & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275-314

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R. & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, 78(4), 1102-1134.

Ainsworth, H., Hewitt, C. E., Higgins, S., Wiggins, A., Torgerson, D. J., & Torgerson, C. J. (2015). Sources of bias in outcome assessment in randomised controlled trials: a case study. *Educational Research and Evaluation*, *21*(1), 3-14.

Amabile, T. M. (1985). Motivation and Creativity: Effects on Motivational Orientation on Creative Writers. *Journal of Personality and Social Psychology*, 48 (2), 393-399.

Amabile, T. M. (2017). *Creativity and Motivation.* YouTube Video. Available at: https://www.youtube.com/watch?v=YRnvox6_o2M (access: 15th June 2018)

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*.

American Philosophical Association (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*: *The Delphi report (Executive Summary).* The California Academic Press.

American Philosophical Association (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*: *The Delphi report Research findings and recommendations prepared for the committee on pre-college philosophy.* ERIC No. ED315423

Anastasi, A. (1988). *Psychological Testing*. 6th edn. London: Collier Macmillan Publishers.

Australian Curriculum, Assessment and Reporting Authority (n.d). *Australian Curriculum: Critical and Creative Thinking*. Available at: https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/critical-and-creative-thinking/ (access: 31st December 2017)

Banks, J. C. R. (1987). *A study of the effects of the critical thinking skills program, Philosophy for Children, on a standardized achievement test.* Ed.D. dissertation. Southern Illinois University.

Bar-Hiller, M. & Attali, Y. (2002). Seek Whence: Answer Sequences and Their Consequences in Key-Balanced Multiple-Choice Tests. The American Statistician, 56 (4), 299-303.

Barnett, R. (1997). *Higher education: A critical business.* Buckingham: The Society for Research into Higher Education and Open University Press.

Barrow, W. (2015). 'I think she's learnt how to sort of let the class speak': Children's perspectives on Philosophy for Children as participatory pedagogy. *Thinking Skills and Creativity*, 17, 76-87.

Bassiri, A & Vaidya, A. J. (2013). Making Everyday Discussions More Philosophical. In *Implementing Philosophy in Elementary Schools: The Washington Elementary School Philosophy Project.* (pp.48-59). Bloomington: Authorhouse.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open ended) account for gender differences in mathematics achievement? *Sex Roles*, 42, 1-21.

Benade, L. (2011). Philosophy for Children (P4C): A New Zealand School-based Action Research Case Study. *New Zealand Journal of Teachers' Work*, 8(2), 141-155.

Benade, L. (2014). Knowledge and Educational Research in the Context of 'Twenty-First Century Learning', *European Educational Research Journal*, 13 (3), 338-349

BERA (2011). Ethical guidelines for educational research. Available at: http://content.yudu.com/Library/A2xnp5/Bera/resources/index.htm?referrerUrl=http://free.yudu.com/item/details/2023387/Bera (access : 20/01/2016)

Berliner, D. C. (2011). The Context for Interpreting PISA Results in the USA: Negativism, Chauvinism, Misunderstanding, and the Potential to Distort the Educational Systems of Nations. In Miguel A. Pereyra, Hans-Georg Kotthoff and Robert Cowen (Ed.) *PISA Under Examination: Changing Knowledge, Changing Tests, and Changing Schools.* (pp. 77-96). Rotterdam: Sense Publishers

Blake, C. (1955). Can history be objective?. *Mind*, 64 (253), 61-78.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, 27(2), 165-174.

Boostrom, R. (1991). The Nature and Functions of Classroom Rules. *Curriculum Inquiry*, 21(2), 193-216.

Brame, C. (2013). *Writing good multiple choice test questions.* Available at: https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/ (access: 5 June 2017)

Brookfield, S. (2012). *Teaching for critical thinking: tools and techniques to help students question their assumptions.* United States: Jossey-Bass.

Brooks, L.A. & Dixon, J. K. (2013). Changing the Rules to Increase Discourse. *Teaching Children Mathematics*, 20(2), 84-89

Bruner, J. (1999). *The Culture of Education*. USA: Harvard University Press.

Burbules, N. C. & Berk, R. (1999). Critical thinking and critical pedagogy: Relations, differences, and limits. In Popkewitz, T. (ed.), *Critical theories in education: Changing terrains of knowledge and politics* (pp.45-65). New York: Routledge.

Burton, S.J., Sudweeks, R.R., Merrill, P.G. & Wood, B. (1991). *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. Brigham Young University Testing Services and The Department of Instructional Science.

Butler, H. A. (2015). Assessing critical thinking in our students. In Wegerif, R., Li, L. & Kaufman, J.C. (ed.) *The Routledge International Handbook of Research on Teaching Thinking* (pp.305-314). London and New York: Routledge

Calsyn, R. J. (2000). A checklist for critiquing treatment fidelity studies. *Mental Health Services Research*, 2(2), 107-113.

Cambridge Assessment (2017). *The Cambridge Approach to Assessment: Principles for designing, administering and evaluating assessment*. Available at: http://www.cambridgeassessment.org.uk/Images/cambridge-approach-to-assessment.pdf (accessed: 22nd June 2017).

Campbell, D.T. & Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-multimethod matrix. *Psychological Bulletin*, 56 (2), 81-105.

Chartered Institute of Educational Assessors (2008). *Oral Language Modifier Guidance Examples.* London. Available at: http://www.batod.org.uk/content/resources/materials/training-materials/language-modification/olm_guidance_examples.pdf (access: 26 November 2016)

Chetty, D. (2014). The elephant in the room: Picturebooks, Philosophy for Children and Racism. *Childhood & philosophy*, 10(19), 11-31.

Child Trends (2016). Educational Attainment: Indicators of Child and Youth Well-being. Available at: https://www.childtrends.org/indicators/educational-attainment/ (access: 15 April 2017)

Coe, R. (1999). Manifesto for Evidence-Based Education. Available at: http://www.cem.org/attachments/ebe/manifesto-for-ebe.pdf (access: 15 April 2017)

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* 2nd edn. New York: Lawrence Erlbaum Associates Publishers.

Cohen, L., Manion, L. & Morrison, K. (2007). *Research Methods in Education.* 6th edn. London: Routledge

Colom, R. Moriyon, F. 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), 50-56.

Cooke, P.A. (2015). *The Impact of Engaging in Philosophy with Middle School Children on the Development of Critical Thinking.* PhD thesis. University of Rochester.

Corazza, G. E. (2016) Potential Originality and Effectiveness: The Dynamic Definition of Creativity, *Creativity Research Journal*, 28 (3), 258-267

Craft, A. (2001). Little c Creativity. In Craft, A., Jeffrey, B. and Leibling, M. (ed.). *Creativity in Education* (pp. 45-61). London: Continuum.

Craft, A. (2005). *Creativity in Schools: Tensions and Dilemmas*. London: Routledge.

Cronbach, L.J. (1961). *Essentials of Psychological Testing.* Harper and Row: New York.

Daniel, M., & Auriac, E. (2011). Philosophy, critical thinking and philosophy for children. *Educational Philosophy and Theory*, 43(5), 415-435.

Davies, M. (2015). A Model of Critical Thinking in Higher Education. In Paulsen, M.B. (ed.), *Higher Education: Handbook of Theory and Research* (pp. 41-92). Springer International Publishing: Switzerland

Davis, G. A. (1999). *Creativity is Forever*. 4th edn. Dubuque, Iowa: Kendall/Hunt Publishing Company.

Demissie, F. (2017). The Praxis of P4C in Teacher Education. In B. Anderson (ed.) *Philosophy for Children: Theories and praxis in teacher education* (pp. 115-121). Oxon: Routledge

DeVellis, R.F.(2006). Classical Test Theory, *Medical Care*, 44(11), S50-S59

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process.* New York: D.C. Health and Company.

Downing, S. M. (2006). Twelve steps for effective test development. In Downing, S.M. & Haladyna, T.M. (ed.), *Handbook of test development* (pp.3-25), London: Lawrence Erlbaum Associates.

Dunlop, L., Compton, K., Clarke, L. & McKelvey- Martin, V. (2015) Child-led enquiry in primary science, *Education 3-13*, 43(5), 462-481. doi:10.1080/03004279.2013.822013.

Edubase2 (n.d.) Available at: http://www.education.gov.uk/edubase/home.xhtml (access: 15 April 2017)

Education Endowment Foundation (2018a). *About.* Available at: https://educationendowmentfoundation.org.uk/about/ (access: 10 January 2018)

Education Endowment Foundation (2018b). Promising Projects. Available at: https://educationendowmentfoundation.org.uk/tools/promising/ (access: 28th July 2018)

El Soufi, N. & See, B. H. (2019). Does explicit teaching of critical thinking improve critical thinking skills of English language learners in higher education? A critical review of causal evidence. *Studies in Educational Evaluation*, 60, 140-162.

Ennis, R. H. (1962). A concept of critical thinking. *Harvard Educational Review*, 32(1), 81-111.

Ennis, R.H. (1964). A definition of critical thinking. *The Reading Teacher*, 17(8), 599-612.

Ennis, R.H. (1984). Problems in Testing Informal Logic, Critical Thinking, Reasoning Ability. *Informal Logic*, 6 (1), 3-9

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational leadership*, 43(2), 44-48.

Ennis, R.H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational researcher*, 18(3), 4-10.

Ennis, R. H. (1996). Critical thinking dispositions: Their nature and assessability. *Informal Logic*, 18(2&3), 165-182.

Ennis, R.H. (2011). *The Nature of Critical Thinking: An Outline of Critical Thinking Dispositions and Abilities.* Revised. Available at: http://faculty.education.illinois.edu/rhennis/documents/TheNatureofCriticalThinking_51711_001.pdf (accessed: 2nd December 2017)

Ennis, R.H. (2015a). *The Nature of Critical Thinking: Outlines of General Critical Thinking Dispositions and Abilities.* Revised. Available at: http://www.criticalthinking.net/longdefinition.html (accessed: 2nd December 2017)

Ennis, R.H. (2015b). *Critical Thinking: A Streamlined Conception*. In Davies, M & Barnett, R. (ed.) *The Palgrave Handbook of Critical Thinking in Higher Education* (pp.31-47). United States: Palgrave Macmillan

Ennis, R.H. & Millman, J. (2005). Cornell Critical Thinking Tests Level X. 5th edn. United States: McNaughton & Gunn, Inc.

Ennis, R.H., Millman, J. & Tomko, T. N. (2005). Cornell Critical Thinking Tests: Administration Manual Level X and Level Z .5th edn, Revised. United States: McNaughton & Gunn, Inc.

Ennis, R.H. & Weir, E. (1985). *The Ennis-Weir Critical Thinking Test: Test, Manual, Criteria, Scoring Sheet. An instrument for teaching and testing*. Pacific Grove: Midwest Publications

Evans, J.St.B.T. (2005). Deductive Reasoning. In Holyoak, K.J. & Morrison, R.G. (ed.) The Cambridge Handbook of Thinking and Reasoning (pp.169-184). United States of America: Cambridge University Press.

Facione, P. A., Sánchez, C. A., Facione, N. C., & Gainen, J. (1995). The disposition toward critical thinking. *The Journal of General Education*, 44(1), 1-25.

Fair, F., Haas, L. E., Gardosik, C., Johnson, D. D., Price, D. P., & Leipnik, O. (2015a). 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), 18-37.

Fair, F., Haas, L. E., Gardosik, C., Johnson, D., Price, D., & Leipnik, O. (2015b). Socrates in the schools: Gains at three-year follow-up. *Journal of Philosophy in Schools*, 2(2), 5-16

Fields, J. I. (1995). Empirical data research into the claims for using philosophy techniques with young children. *Early Child Development and Care*, 107 (1), 115-128

Fisher, A. (2010). *Critical Thinking: An Introduction.* Cambridge: Cambridge University Press.

Fisher, A. & Scriven, M. (1997). *Critical Thinking: Its Definition and Assessment.* Norwich: Centre for Research in Critical Thinking.

Fisher, R. (2003). *Teaching Thinking*. 2nd edn. London and New York: Continuum.

Fisher, R. (2005). *Teaching children to think*. 2nd edn. United Kingdom: Nelson Thornes

Fung, D. (2014). Promoting critical thinking through effective group work: A teaching intervention for Hong Kong primary school students. *International Journal of Educational Research*, 66, 45-62.

Gallagher, D. & Grimm, L. R. (2018). Making an impact: The effects of game making on creativity and spatial processing. *Thinking Skills and Creativity*, 28, 138-149.

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

Gasparatou, R. & Ergazaki, M. (2015). Students' Views about their Participation in a Philosophy Program. *Creative Education*, 6, 726-237.

Gasparatou, R. & Kampeza, M. (2012) Introducing P4C in Kindergarten in Greece. *Analytic Teaching and Philosophical Praxis*, 33(1), 72-82

Gazzard, A. (2012). Do you Need to Know Philosophy to Teach Philosophy to Children? A Comparison of Two Approaches, *Analytic Teaching and Philosophical Praxis*, 33(1), 45-53

Getzels, J.W. & Jackson, P. W. (1962). *Creativity and Intelligence: Explorations with Gifted Students*. London and New York: John Wiley and Sons Inc.

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), 63-89.

Golding, C. (2007) Pragmatism, Constructivism and Socratic Objectivity: The pragmatist epistemic aim of philosophy for children in: Creativity, Enterprise and Policy—New Directions in Education, *36th Annual Philosophy of Education Society of Australasia Conference*, 6–9 December, Wellington

Gorard, S. (2001). *Qualitative Methods in Educational Research: The role of Numbers Made Easy* (2nd edn). London: Continuum.

Gorard, S. (2013). Research Design: Creating Robust Approaches for the Social Sciences. Thousand Oaks, CA: SAGE

Gorard, S. (2015a). Rethinking 'quantitative' methods and the development of new researchers. *Review of Education*, 3 (1), 72-96. doi: 10.1002/rev3.3041

Gorard, S. (2015b). A proposal of judging the trustworthiness of research findings. researchED Magazine. Available at: http://www.workingoutwhatworks.com/en-GB/Magazine/2015/1/Trustworthiness_of_research (access: 19th February 2017)

Gorard, S. (2016). Damaging real lives through obstinacy: re-emphasising why significance testing is wrong. *Sociological Research Online*, 21(1), 1-14. doi: 10.5153/sro.3857

Gorard, S. & Gorard, J. (2016). What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding, *International Journal of Social Research Methodology*, 19(4), 481-490

Gorard, S., See, B.H. & Morris, R. (2016). *The most effective approaches to teaching in primary schools: Rigorous evidence on effective teaching*. Saarbrücken: LAP LAMBERT Academic Publishing.

Gorard, S., See, B.H. & Siddiqui, N. (2017) *The Trials of Evidence-Based Education: The Promises, Opportunities and Problems of Trials in Education*. London: Routledge.

Gorard, S., Siddiqui, N. & See, B.H. (2015). Philosophy for Children: Evaluation Report and Executive Summary. Available at https://educationendowmentfoundation.org.uk/uploads/pdf/Philosophy_for_Children.pdf (accessed: 1 November 2015)

Gorard, S., Siddiqui, N., & See, B. H. (2017). Can 'Philosophy for children 'improve primary school attainment?. *Journal of Philosophy of Education*, 51(1), 5-22.

GOV.UK (n.d.). *Compare school and college performance*. Available at: https://www.compare-school-performance.service.gov.uk/ (access: 15 April 2017)

Green, L. & Condy, J. (2016). Philosophical enquiry as a pedagogical tool to implement the CAPS curriculum: Final year pre-service teachers' perceptions. *South African Journal of Education*, 36 (1), 1-8.

Gronlund, N. E. (1982). Planning the test. In *Constructing Achievement Tests* (pp.18-35). Englewood Cliffs, London: Prentice-Hall Inc.

Guilford, J.P. (1950). Creativity. *American Psychologist*, 5(9), 444 - 454.

Guilford, J.P. (1956). The Structure of Intellect. *Psychological Bulletin*, 53 (4), 267-293.

Guilford, J.P. (1967). *The nature of Human Intelligence*. Mc Graw-Hill Book Company

Gupta, S. K. (2011). Intention-to-treat concept: a review. *Perspectives in clinical research*, 2(3), 109- 112. Doi: 10.4103/2229-3485.83221

Haladyna, T. M. (1994). Writing the test item. In *Developing and Validating Multiple-Choice Test Items* (pp. 61-86), Hillsdale, N.J: LEA

Haladyna, T.M., Downing, S.M. & Rodriguez, M.C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 15(3), 309–334

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist*, 53(4), 449–455.

Halpern, D.F. (2010). *Halpern Critical Thinking Assessment*. Vienna: Schuhfried.

Harrison-Barbet, A. (2001). *Mastering Philosophy*. 2nd edn. Basingstoke: Palgrave Macmillan

Haynes, J. (2008). *Children as philosophers*. London: Routledge.

Haynes, J. & Murris, K. (2012) *Picturebooks, Pedagogy and Philosophy*. London: Routledge.

Hedayati, M. & Ghaedi, Y. (2009). Effects of Philosophy for Children through the Community of Inquiry method on the improvement of interpersonal relationship skills in primary school students. *childhood & philosophy*, 5(9), 199-217

Hewitt, M.A. & Homan, S.P. (2003). Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction*, 43(2), 1-16

Higgins, S. (2017). *The EEF Toolkit has revealed the academic garden to teachers.* [Blog Post]. Available at: https://www.tes.com/news/school-news/breaking-views/eef-toolkit-has-revealed-academic-secret-garden-teachers (access: 10 January 2018)

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A.B. , Coleman,R., Henderson, P., Major, L.E., Coe, R. & Mason, D. (2016) *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit*. Available at: https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit (access: 10 January 2018)

Hirsch, E. D. (2001). Seeking breadth and depth in the curriculum. *Educational Leadership*, 59(2), 22-25.

Hirsch, E. D. (2011). Beyond Comprehension: We Have yet to Adopt a Common Core Curriculum that Builds Knowledge Grade by Grade - But We Need to. *American Educator*, 34(4), 30-36.

Hughes, D. J. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In Irwing, P., Booth, T. & Hughes,D.J. (ed.). *Wiley handbook of psychometric testing: A multidisciplinary approach to survey, scale, and test development*. (pp. 751-779). Hoboken, NJ: Wiley.

Ioannou, S., Chatziefraimidou, A. & Ventista, O.M. (under review). Teachers' Perceptions Concerning the Teaching of Ethics in Primary School Classrooms in Greece and Cyprus: Teaching Approaches and Resources.

Ioannou, S., Georgiou, K. & Ventista, O.M. (2017) Teaching Philosophy through Paintings: A Museum Workshop. *Analytic Teaching and Philosophical Praxis*, 38 (1), 62-83

Jahani, R., & Akbari, A. The Effect of P4C (Process & Content Approach) on Creativity of the Six Grade Students (Case Study: Sixth Grade Male Students of Jolge Rokh Area). *Scinzer Journal of Accounting and Management*, 2 (3), 10-15. doi: 10.21634/SJAM.2.3.1015

Jahani, R., Nodehi, H., & Akbari, A. (2016). Effect Of The P4C (Philosophy For Children As A Content Approach) On Moral Gudgment Of sixth Grade Students (Case Study: Jolgeh Rokh Area). *Scinzer Journal of Humanities*, 2(1), 19-23.

James, K. & Taylor, A. (2012). Positive Creativity and Negative Creativity (and Unintended Consequences). In Cropley, D.H., Cropley, A.J., Kaufman, J.C. and Runco, M.A. (eds.) *The Dark Side of Creativity.* 4th edn. (pp. 33-56) Cambridge: Cambridge University Press

Jenkins, J. (1986). Philosophy for Children Programme at a Gloucestershire Comprehensive School in Great Britain. *Thinking: The Journal of Philosophy for Children*, 6 (3), 34-37

Jenkins, J.S. (1995). Mozart and medicine in the eighteenth century. *Journal of the Royal Society of Medicine*, 88, 408P-413P

Jenkins. P. & Lyle, S. (2010). Enacting dialogue: the impact of promoting Philosophy for Children on the literate thinking of identified poor readers, aged 10, *Language and Education*, 24(6), 459-472

Jo, S. H. (2001). Literacy: Constructing meaning through philosophical inquiry. *Analytic Teaching*, 21(1), 44-52.

Jones, B.F. & Idol, L. (1990). Introduction. In Jones, B.F. and Idol, L. (ed) *Dimensions of Thinking and Cognitive Instruction* (pp. 1-13) Hove and London: Lawrence Erlbaum Associates Publishers.

Kane, M. T. (1990). *An Argument-based approach to Validation*. Research Report. Iowa: American College Testing Programme.

Kane, M.T. (1992). An Argument-based approach to Validity. *Psychological Bulletin*, 112(3), 527-535.

Karadağ, F. & Demirtaş, V. Y. (2018). The Effectiveness of The Philosophy with Children Curriculum on Critical Thinking Skills of Pre-School Children. *Education & Science/Egitim ve Bilim*, 43(195), 19-40.

Kaufman, J.C. (2006). Self-Reported Differences in Creativity by Ethnicity and Gender. *Applied Cognitive Psychology*, 20, 1065-1082.

Kaufman, J.C. & Sternberg, R.J. (2007). Creativity, *Change*, 39 (4), 55-58

Kellner, R., & Benedek, M. (2017). The role of creative potential and intelligence for humor production. *Psychology of Aesthetics, Creativity, and the Arts*, 11(1), 52-58.

Kennedy, D. (2004). The Philosopher as Teacher: The role of a Facilitator in a Community of Philosophical Inquiry. *Metaphilosophy*, 35(5), 744-765.

Kim, K.H. (2006). Can we trust Creativity Tests? A Review of the Torrance Test of Creative Thinking (TTCT). *Creativity Research Journal*,18(1), 3-14. Doi: 10.1207/ s15326934crj1801_2

Kong, L. N., Qin, B., Zhou, Y. Q., Mou, S. Y. & Gao, H. M. (2014). The effectiveness of problem-based learning on development of nursing students' critical thinking: A systematic review and meta-analysis. *International journal of nursing studies*, 51(3), 458-469.

Koretz, D. (2006). *Measuring Up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Kuha, J. & Sturgis, P. (2016) Comment on 'What to do instead of significance testing? Calculating the "number of counterfactual cases needed to disturb a finding'. by Stephen Gorard and Jonathan Gorard. *International Journal of Social Research Methodology*, 19 (4), 491-495.

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

LeCompte, M. (1978). Learning to Work: The Hidden Curriculum of the Classroom. *Anthropology & Education Quarterly*, 9 (1), 22-37.

Lipman, M. (1976). Philosophy for children. *Metaphilosophy*, 7 (1), 17-39.

Lipman, M. (1982). Philosophy for children. *Thinking: The Journal of Philosophy for Children*, 3(3/4), 35-44.

Lipman, M. (1985). Philosophy for Children and Critical Thinking. *National Forum*, 65(1), 18-21

Lipman, M. (1987). Critical thinking: What can it be?, *Analytic Teaching*, 8(1), 5-12.

Lipman, M. (1992). On Writing a Philosophical Novel. In A.M. Sharp & R.F. Reed (eds.) *Studies in Philosophy for Children: Harry Stottlemeier's Discovery* (pp.3-7). Philadelphia: Temple University Press.

Lipman, M. (1998). Teaching students to think reasonably: Some findings of the Philosophy for Children program. *The Clearing House*, 71(5), 277-280.

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

Lipman, M. (2009). Philosophy for Children: Some Assumptions and Implications. In Eva Marsal, Takara Dobashi and Barbara Weber (eds.), *Children Philosophize Worldwide*. Frankfurt: Peter Lang.

Lyle, S. (2017). The construct of the child: the 'C' in PwC. In B. Anderson (ed.) *Philosophy for Children: Theories and praxis in teacher education* (pp. 25-36). Oxon: Routledge

Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment*, 1(1), 1-11.

Marashi, S. M. (2008). Teaching philosophy to children: A new experience in Iran. *Analytic Teaching*, *27*(1), 12-15.

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: evidence from 31 countries. *Oxford Review of Education*, 34(1), 89-109.

Mascitelli-Morey, S. (2013). Assessing the Effectiveness of Classroom Sessions. In *Implementing Philosophy in Elementary Schools: The Washington Elementary School Philosophy Project*. (pp. 69-77). Bloomington: Authorhouse.

Matthews, G. B. (1978). Are Children Philosophical? In M. Lipman & A.M. Sharp (eds.) *Growing Up with Philosophy* (pp.63-77). Philadelphia: Temple University Press.

Matthews, G.B. (1984). *Dialogues with Children.* London: Harvard University Press.

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

McCall, C.C. (2009). *Transforming Thinking: Philosophical Inquiry in the primary and secondary classroom.* Oxon: Routledge.

McPeck, J. E. (1981). *Critical Thinking and Education.* Oxford: Martin Robertson.

McPeck, J. E. (1985). Paul's critique of Critical thinking and education. *Informal Logic*, 7(1), 45-54.

McPeck, J.E. (1990). Critical thinking and subject specificity: A reply to Ennis. *Educational Researcher*, 19(4), 10-12.

Meir, S. & McCann, J. (2017). An evaluation of P4C. In B. Anderson (ed.) *Philosophy for Children: Theories and praxis in teacher education* (pp.83-92). Oxon: Routledge

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, 50(9), 741-749.

Mitra, S. (2000). *Children and the Internet: New Paradigms for Development in the 21st Century*. Asian Science and Technology Conference. Tokyo.

Morante, E. A. & Ulesky, A. (1984). Assessment of Reasoning abilities. *Educational Leadership*, 42(1), 71-74.

Morris, S.B. (2008). Estimating Effect Sizes for Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, 11 (2), 364-386.

Murris, K. (1992). *Teaching philosophy with picturebooks*. London: Infonet Publications.

Neill, A. S. (1960). Summerhill: A Radical Approach to Child Rearing. New York: Hart Publishing Company.

Newton, D.P.(2014). *Thinking with Feeling: Fostering productive thought in the classroom*.New York: Routledge

Newton, D. P. (2015). There's more to thinking than the intellect. In Wegerif,R., Li, L. & Kaufman, J.C. (ed.) *The Routledge International Handbook of Research on Teaching Thinking* (pp.58-68). London and New York: Routledge

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, 14(2), 149-170.

Nia, A.T. (2014a). Investigate the Effect the Philosophy for Children Program (p4c) on Reducing Trait Anger in Teens. *Journal of Educational and Management Studies*, 4 (2), 449-455

Nia, A.T. (2014b). Foster Self-Esteem in Adolescents: Lipmann Approach. *Journal of Educational and Management Studies*, 4(2), 391-396.

Niu, L., Behar-Horenstein, L. S. & Garvan, C. W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educational Research Review*, 9, 114-128.

Norris, S.P. & Ennis, R.H. (1989). *Evaluating Critical Thinking.* Pacific Grove, CA: Midwest Publications

Norris, S. P. & King, R. (1984). *The design of a Critical Thinking Test on Appraising Observations. Studies in Critical Thinking*. Research Report No1. Canada: Institute for Educational Research and Development, Memorial University of Newfoundland

Nusbaum, E. C., Silvia, P. J. & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. *Psychology of Aesthetics, Creativity, and the Arts*, 11(2), 231-241.

Office for National Statistics (n.d.). *Gender Identity*. Available at: https://www.ons.gov.uk/methodology/classificationsandstandards/measuringequality/genderidentity (access: 15th April 2017)

Ofsted (2018). *School Inspection Handbook.* Available at: https://www.gov.uk/government/publications/school-inspection-handbook-from-september-2015 (access: 5 November 2018)

Paul, R. (1985). McPeck's mistakes. *Informal Logic*, 7(1), 35-43.

Paul, R. W. (1993). The logic of creative and critical thinking. *American Behavioral Scientist*, 37(1), 21-39.

Paul, R. & Elder, L. (2005). *A Guide for Educators to Critical Thinking Competency Standards: Standards, Principles, Performance Indicators, and Outcomes with a Critical Thinking Master Rubric.* Foundation for Critical Thinking.

Paul, R., & Elder, L. (2008). *The miniature guide to critical thinking: Concepts & tools*. 5th edn. Foundation Critical for Critical Thinking.

Piaget, J. (1999). *The Construction of Reality in the Child*. Oxon: Routledge

Piirto, J. (2010). The five core attitudes, seven I's, and general concepts of the creative process. In Beghetto, R. A. and Kaufman, J. C. (eds.) *Nurturing creativity in the classroom* (pp.142-171). New York: Cambridge University Press.

Piirto, J. & Ford, R. (2000). The Piirto pyramid of talent development. *Gifted Child Today*, 23(6), 22-29.

Plucker, J. A., Baghetto, R.A. & Dow, G.T. (2004). *Why Isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research*, Educational Psychologist, 39(2), 83-96. doi: 10.1207/s15326985ep3902_1

Plucker, J. A. & Makel, M. (2010). Assessment of Creativity. In Kaufman J.C. and Sternberg, R.J. (eds.) *The Cambridge Handbook of Creativity* (pp.48-73). United States of America: Cambridge University Press.

Plucker, J. A. & Renzulli, J.S. (1999). Psychometric Approaches to the study of Human Creativity. In Sternberg, R.J. (ed.) *Handbook of Creativity* (pp. 35-61).New York: Cambridge University Press.

Pourtaghi, V., Hosseini, A. & Hejazi, E. (2014). Effectiveness of implementing philosophy for children program on students' creativity. *Scientific Journal of Pure and Applied Sciences, 3*(6), 375-380.

Pring, R. (1980). *Knowledge and Schooling*. Somerset, England: Open Books

Pring. R. (2007). *John Dewey: A Philosopher of Education for our time?* London: Bloomsbury.

Pritchard, M. S. (1992). Critical Thinking: Problem Solving or Problem Creating? In Sharp, A.M. and Reed, R.F. (eds) *Studies in Philosophy for Children: Harry Stottlemeier's Discovery* (pp.87-95). Philadelphia: Temple University Press.

Rahdar, A., Pourghaz, A., & Marziyeh, A. (2018). The Impact of Teaching Philosophy for Children on Critical Openness and Reflective Skepticism in Developing Critical Thinking and Self-Efficacy. *International Journal of Instruction*, 11(3), 539-556

Reardon, S.F., Kalogrides, D., Fahle, E.M., Podolsky, A & Zárate, R.C. (2018). The Relationship Between Test Item Format and Gender Achievement Gaps on Math and ELA Tests in Fourth and Eighth Grades. *Educational Researcher*, 47 (5).

Reed-Sandoval, A. & Sykes, A. C. (2017). Who talks? Who listens? Taking 'positionality' seriously in Philosophy for Children. In Gregory M.R., Haynes, J. & Muris, K. (ed.) *The Routledge International Handbook of Philosophy for Children.* (pp. 219-226). Oxon: Routledge

Reznitskaya, A. & Glina, M. (2013). Comparing Student Experiences with Story Discussions in Dialogic Versus Traditional Settings, *The Journal of Educational Research*, 106(1), 49-63, DOI: 10.1080/00220671.2012.658458.

Reznitskaya, A., Glina, M., Carolan, B., Michaud, O., Rogers, J., & Sequeira, L. (2012). Examining transfer effects from dialogic discussions to new tasks and contexts. *Contemporary Educational Psychology*, 37(4), 288-306.

Reznitskaya, A. & Wilkinson, I. A. (2017). Truth matters: Teaching young students to search for the most reasonable answer. *Phi Delta Kappan*, 99(4), 33-38.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13.

Rosenthal, R. (1991). *Meta-analytic procedures for social research: Revised edition*. London: Sage Publications.

Runco, M.A. (2004). Creativity. *Annual Review of Psychology*, 55, 657-687.

Runco, M.A. (2012). Creativity has no Dark Side. In Cropley, D.H., Cropley, A.J., Kaufman, J.C. and Runco, M.A. (eds.) *The Dark Side of Creativity.* 4th edn. (pp. 15-32) Cambridge: Cambridge University Press

Runco, M. A. & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, 24(1), 92-96.

Russell, R.L. (1988). Children's Philosophical Inquiry into Defining Art: A Quasi-Experimental study of Aesthetics in the Elementary Classroom. *Studies in Art Education*,29(3), 282-291.

Said-Metwaly, S., Van den Noortgate, W. & Kyndt, E. (2017). Approaches to Measuring Creativity: A Systematic Literature Review. *Creativity: Theories–Research-Applications*, 4(2), 238-275.

Sala, G. & Gobet, F. (2019). Cognitive Training Does Not Enhance General Cognition. *Trends in cognitive sciences*, 23 (1), 9-20.

Santos, S., Jiménez, S., Sampaio, J. & Leite, N. (2017). Effects of the Skills4Genius sports-based training program in creative behavior. *PloS one*, 12(2), e0172520.

Sanz de Acedo Lizarraga, L.M., Ugarte, D.M., Iriarte, D. Sanz de Acedo Baquedano, T. (2003). Immediate and long-term effects of a cognitive intervention on intelligence, self-regulation, and academic achievement. *European Journal of Psychology of Education*, 18(1), 59-74.

SAPERE: Philosophy for Children, Colleges, Communities (2015a). What is P4C? Available at: http://www.sapere.org.uk/default.aspx?tabid=162 (access: 15 April 2017)

SAPERE: Philosophy for Children, Colleges, Communities (2015b). About SAPERE. Available at: http://www.sapere.org.uk/Default.aspx?tabid=70 (access: 23 September 2016)

SAPERE (2015c). Community of Enquiry. Available at https://www.sapere.org.uk/Default.aspx?tabid=76 (access: 6 November 2018)

Säre, E., Luik, P., & Tulviste, T. (2016). Improving Pre-schoolers' reasoning skills using the philosophy for children programme. *Trames: A Journal of the Humanities and Social Sciences*, 20(3), 273-295.

Saretsky, G. (1972). The OEO PC experiment and the John Henry effect. *The Phi Delta Kappan*, *53*(9), 579-581.

Sargent, C., Byrne, A., O'Donnell, S. & White, E. (2010). *Thematic Probe: Curriculum Review in the INCA countries: June 2010.* NFER. Available at: http://webarchive.nationalarchives.gov.uk/20130220111913/http://www.inca.org.uk/Curriculum_review_probe_final_01_dec_2010.pdf (access: 21 October 2018)

Sasseville, M. (1994). Self-esteem, logical skills and philosophy for children, *Thinking*, 4(2), 30–32.

Schleifer, M., Daniel, M. F., Peyronnet, E., & Lecomte, S. (2003). The Impact of Philosophical Discussions on Moral Autonomy, Judgment, Empathy and the Recognition of Emotion in Five Tear Olds. *Thinking: The Journal of Philosophy for Children*, 16(4), 4-12.

Schleifer, M., & Poirier, G. (1996). The effect of philosophical discussions in the classroom on respect for others and non-stereotypic attitudes. *Thinking: The Journal of Philosophy for Children*, 12(4), 32-34.

See, B. H., Gorard, S., & Siddiqui, N. (2017). Can explicit teaching of knowledge improve reading attainment? An evaluation of the Core Knowledge curriculum. *British Educational Research Journal*, 43(2), 372-393.

See, B. H. & Kokotsaki, D. (2016). Impact of arts education on children's learning and wider outcomes. *Review of Education*, 4(3), 234-262.

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-experimental designs for generalised causal inference*, Boston : Houghton Mifflin

Shaheen, R. (2010). An investigation into the factors enhancing or inhibiting primary school children's creativity in Pakistan. Available at: http://etheses.bham.ac.uk/1239/1/Shaheen10PhD.pdf (access: 29 February 2016)

Shatalebi, A., & Hedayati, M. (2016). Investigating the Effects of "Philosophy for Children" Program on the Reduction of Psychosomatic Disorders Symptoms in 9-11 Age Boys. *International Letters of Social and Humanistic Sciences*, 66, 1-9.

Sherman, L. W., Gottfredson, D.C., MacKenzie, D. L., Eck, J., Reuter, P. & Bushway, S.D. (1997). *Preventing Crime: What Works, What Doesn't, What's Promising*. Washington, DC: National Institute of Justice

Shipman, V. (1983). *New Jersey Test of Reasoning Skills*. Upper Montclair, NJ: Montclair State College.

Sick, J. (2008). Rasch Measurement in Language Education: Part 1. *JALT Testing & Evaluation SIG Newsletter*, 12(1), 1-6

Siddiqui, N., Gorard, S. & See, B.H. (2017). *Non-cognitive impacts of Philosophy for Children programme*. Durham University: Durham. Available at Nuffield Foundation website: http://www.nuffieldfoundation.org/non-cognitive-impacts-philosophy-children (accessed: 5th March 2017)

Siddiqui, N. & Ventista, O.M. (2018). A review of school-based interventions for the improvement of social emotional skills and wider outcomes of education. *International Journal of Educational Research*, 90, 117-132.

Siegel, H. (1988). *Educating reason: Rationality, critical thinking and education*. London: Routledge.

Simpson, A. (2017). The misdirection of public policy: comparing and combining standardised effect sizes. Journal of Education Policy. doi: 10.1080/02680939.2017.1280183

Slade, C. (1989). Logic in the classroom. *Thinking*, 8 (2), 14-20

Slavin, R. E. & Smith, D. (2008). Effects of Sample Size on Effect Size in Systematic Reviews in Education. *Annual meeting of the Society for Research on Effective Education*, 2-4 March, Crystal City, Virginia.

Smith, E. (2017). Secondary data. In Coe, R., Waring, M., Hedges, L.V. & Arthur, J. (ed). *Research Methods & Methodologies in Education*. 2nd edn. (pp. 122-129). London: Sage.

Snape, D., Manclossi, S., Hassell, C., Osborn, E., Martin, G., Sidney, I., Pyle, E. & Cochrane, A. (2018). Children's and young people's experiences of loneliness: 2018. Office for National Statistics. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/childrensandyoungpeoplesexperiencesofloneliness/2018 (access: 9 December 2018)

Spielman, A. (2018). *HMCI commentary: curriculum and the new education inspection framework.* Available at: https://www.gov.uk/government/speeches/hmci-commentary-curriculum-and-the-new-education-inspection-framework (access: 5 November 2018)

Splitter, L. J. (1992). A Guided Tour of the Logic in Harry Stottlemeier's Discover. In Sharp, A.M. and Reed, R.F. (eds) *Studies in Philosophy for Children: Harry Stottlemeier's Discovery* (pp. 107-124). Philadelphia: Temple University Press.

Sprod, T. (1998). “I can change your opinion on that”: Social constructivist whole class discussions and their effect on scientific reasoning. *Research in Science Education*, 28(4), 463-480.

Statistics New Zealand. (2014). *Gender identity: Developing a statistical standard.*

Statistics New Zealand (2015). *Statistical standard for gender identity*.

Sternberg, R. J. (1986). *Critical Thinking: Its Nature, Measurement, and Improvement.* (Eric Document Reproduction No. 272882).

Sternberg, R. J. (2012). The Dark Side of Creativity and How to Combat it. In Cropley, D.H., Cropley, A.J., Kaufman, J.C. and Runco, M.A. (eds.) *The Dark Side of Creativity.* 4th edn. (pp. 316-328) Cambridge: Cambridge University Press

Sternberg, R.J. & Lubart, T. I. (1999). The Concept of Creativity: Prospects and Paradigms. In Sternberg, R.J. (ed.) *Handbook of Creativity* (pp. 3-15).New York: Cambridge University Press.

Sternberg, R.J., Lubart, T.I., Kaufman, J.C. & Pretz, J.E. (2005). Creativity. In Holyoak, K.J. & Morrison, R.G. (ed.) *The Cambridge Handbook of Thinking and Reasoning* (pp.351-369). United States of America: Cambridge University Press.

Stevens, S.S. (1935). The Operational Basis of Psychology. The *American Journal of Psychology*, 47 (2), 323-330.

Stokell, K., Swift, D. & Anderson, B. (2017). P4C in the primary school. In B. Anderson (ed.) *Philosophy for Children: Theories and praxis in teacher education* (pp. 66-71). Oxon: Routledge

Storer, T. (2018). *The Effect of Project Based Learning on the Creativity of Elementary Students.* PhD Thesis. Wilkes University.

Swain, J., Cara, O. & Litster, J. (2014). *Doing Philosophy in schools: Final Report.* Unpublished report.

Sutcliffe, R. (2017). The evolution of Philosophy for Children in the UK. In B. Anderson (ed.) *Philosophy for Children: Theories and praxis in teacher education* (pp.3-13). Oxon: Routledge

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical education*, 42(2), 198-206.

Tarrant, M., Ware, J. & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, 9(1), 40.

Thomas, J. C. (1992). The development of Reasoning in Children through Community of Inquiry. In Sharp, A.M. and Reed, R.F. (eds) *Studies in Philosophy for Children: Harry Stottlemeier’s Discovery* (pp. 96-104). Philadelphia: Temple University Press.

Tian, S. & Liao, P. F. (2016). Philosophy for children with learners of English as a foreign language. *Journal of Philosophy in Schools*, 3(1), 40-58.

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies*, 4(1), 1-17.

Torgerson, D.J. & Torgerson, C.J. (2008). *Designing Randomised trials in health, education and social sciences: an introduction*. New York: Palgrave Macmillan.

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. (2018). Implementation fidelity in computerised assessment of book reading. *Computers & Education*, 116, 176-190.

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

Torrance, P.E. (1962). *Guiding Creative Talent*. Englewood Cliffs, N.J.: Prentice-Hall Inc.

Torrance, P.E. (1988). The nature of creativity as manifest in its testing. In Sternberg, R. J. (ed.) *The Nature of Creativity: Contemporary Psychological Perspectives* (pp. 43 - 75). Cambridge, UK: Cambridge University Press.

Torrance, E. P., Ball, O. E. & Safter H.T. (2008). *Torrance Tests of Creative Thinking: Streamlined Scoring Guide for Figural Forms A and B.* Bensenville: Scholastic Testing Service Inc.

Traub, R.E. & Rowley, G.L. (1991). *An NCME Instructional Module on Understanding Reliability.* Available at National Council on Measurement in Education website. Instructional Topics in Educational Measurement Series (ITEMS): https://members.ncme.org/ncme/NCME/NCME/Publication/ITEMS.aspx (access: 21 July 2018).

Treharne, G. J., & Beres, M. A. (2016). Writing survey questions to operationalise sex, gender identity, and sexual orientation in new Zealand: Perspectives from psychological and sociological research with the LGBTQ community. *New Zealand Sociology*, *31*(1), 173-180.

Trickey, S. & Topping, K.J. (2004). 'Philosophy for children': a systematic review, *Research Papers in Education*, 19(3), 365-380, doi: 10.1080/0267152042000248016

Trickey, S. & Topping, K.J. (2006). Collaborative Philosophical Enquiry for School Children: Socio-Emotional Effects at 11-12 Years. *School Psychology International*, 27(5), 599-614

Vally, Z., Salloum, L., AlQedra, D., El Shazly, S., Albloshi, M., Alsheraifi, S. & Alkaabi, A. (2019). Examining the effects of creativity training on creative production, creative self-efficacy, and neuro-executive functioning. *Thinking Skills and Creativity*, 31, 70-78.

Vansieleghem, N. (2011). Philosophy with children as an exercise in parrhesia: An account of a philosophical experiment with children in Cambodia. *Journal of philosophy of Education*, 45(2), 321-337.

Vansieleghem, N. & Kennedy, D. (2011). What is Philosophy for Children, What is Philosophy with Children - After Matthew Lipman?. *Journal of Philosophy of Education*, 45(2), 171-182.

Ventista, O.M. (2017). *Multiple-Choice Items: A guide for teachers*. Evidence Based Education.

Ventista, O. M. (2018a). Multi-Trait Multi-Method Matrices for the Validation of Creativity and Critical Thinking Assessments for Secondary School Students in

England and Greece. *International Journal of Assessment Tools in Education*, 5(1), 15-32.

Ventista, O.M. (2018b). A Literature Review of Empirical Evidence on the Effectiveness of Philosophy for Children. In Duthie, E., Garcia, F.M. & Robles, R. (Ed.). *Parecidos de familia. Propuestas actuales en Filosofía para Niños / Family ressemblances.Current proposals in Philosophy for Children*. (pp. 448-469). Madrid: Anaya.

Ventista, O.M. & Coe, R. (2015). Can Creativity and Critical Thinking be assessed as general constructs or as subject-specific skills?, T*he European Conference of Educational Research (ECER): Education and Transition*, 8-11th September 2015, Budapest, Hungary

Ventista, O.M. & Paparoussi, M. (2014). Discussing the 'missing piece' and fulfilment: philosophical discussion in the primary school with literature as stimulus [In Greek], Κείμενα, 20, 1-13. Available at: http://keimena.ece.uth.gr/main/index.php?option=com_content&view=article&id=319:20-01&catid=63:tefxos20&Itemid=100

Ventista, O.M. & Paparoussi, M. (2015). Discussing the 'missing piece' and fulfilment: philosophical discussion in the primary school with literature as stimulus. *The European Conference of Educational Research (ECER): Education and Transition*. 7-11th September, Budapest, Hungary

Ventista, O. M. & Paparoussi, M. (2016). Introducing a philosophical discussion in your classroom: an example of a community of enquiry in a Greek primary school. *Childhood & philosophy*, 12(25), 611-629.

Wallas, G. (1926). *The Art of Thought*. London: Butler & Tanner Ltd.

Wartenberg, T. E. (2009). *Big Ideas for Little Kids.* Plymouth, UK: Rowman & Littlefield Publishers, Inc.

Wartenberg, T.E. (2014). Assessing an Elementary School Philosophy Program. *Thinking: The Journal of Philosophy for Children*, 20 (3-4), 90-94

Watson, G. & Glaser, E. (2002). *Watson- Glaser Critical Thinking Appraisal UK Edition. Practice Test.* England: Pearson Assessment. Available at: http://www.pearsonvue.com/phnro/wg_practice.pdf (access: 8 March 2016)

Weber, B. &Wolf, A. (2017). Questioning the Question: A hermeneutical perspective on the 'art of questioning' in a community of philosophical enquiry. In In Gregory M.R., Haynes, J. & Muris, K. (ed.) *The Routledge International Handbook of Philosophy for Children.* (pp. 74-82). Oxon: Routledge

Weisberg, R. W. (2015). On the usefulness of "value" in the definition of creativity. *Creativity Research Journal*, 27(2), 111-124.

Williams, S. (1993). *Evaluating the Effects of Philosophical Enquiry in a Secondary School.* Derbyshire, England: Derbyshire County Council

Wilson, M. (2005) *Constructing Measures: An item Response Modeling Approach.* NJ: Lawrence Erlbaum Associates Inc.

Wyse, D. & Ferrari, A. (2015). Creativity and education: Comparing the national curricula of the states of the European Union and the United Kingdom. *British Educational Research Journal*, 41(1), 30-47. doi: 10.1002/berj.3135

Yan, S. (2017). *Meta-Analysis of the Effects of Philosophy for Children Program on Students' Cognitive Outcome.* Unpublished PhD Thesis. Available at: https://oaktrust.library.tamu.edu/handle/1969.1/161473 (access 11 August 2018)

Yip, D. Y., Chiu, M. M., & Ho, E. S. C. (2004). Hong Kong student achievement in OECD-PISA study: Gender differences in science content, literacy skills, and test item formats. *International Journal of Science and Mathematics Education*, 2(1), 91-106

Youssef, C., Campbell, M., & Tangen, D. (2016). The Effects of Participation in a P4C Program on Australian Elementary School Students. *Analytic Teaching and Philosophical Praxis*, 37(1), 1-19.

Zimmaro, D.M. (2016). *Writing Good Multiple-choice Exams.* University of Texas: *Center* for Teaching and Learning