Investigating the Effect of Students' Personality Traits Towards Improving Pair Programming's Effectiveness as a Pedagogical Tool for CS/SE Education

Norsaremah Salleh

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science, The University of Auckland, 2010.

This thesis is for examination purposes only and may not be consulted or referred to by any persons other than those involved in the examination process.

Abstract

Pair programming (PP) is a practice where two people sit side by side, using only one computer, and work collaboratively on the same design, algorithm, code or test. Given that each member of a pair presents their own personality traits, numerous studies have investigated the possible effect that these traits may have upon the pair's work. However, the results of a Systematic Literature Review carried out as part of this research showed that despite existing evidence of PP's effectiveness in a higher education context, previous PP studies presented inconsistent results in terms of the effect or influence of personality towards that effectiveness. In addition, the personality instrument that had been previously used was also criticized by personality psychologists for being unreliable in measuring an individual's personality traits.

The aim of this research was to improve the implementation of PP as a pedagogical tool for use in Higher Education through understanding the impact that the variation in the personality composition of paired students has towards their academic performance. The personality measurement framework used in this research was the Five-Factor Model, which comprises five broad traits (O*penness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism*). We investigated the effects of personality composition on PP's effectiveness by conducting a series of formal experiments at the University of Auckland. We used as our subjects undergraduate students attending either an introductory programming course or an intermediate software design and construction course. This thesis reports the five formal experiments that separately investigated three of the five traits part of the Five-Factor Model, namely Conscientiousness, Neuroticism, and Openness to experience. These three traits were selected because evidence shows that they are educationally important and relevant for higher education.

Our findings showed that two of the three personality traits investigated – Conscientiousness and Neuroticism, did not present a statistically significantly effect upon paired students´ academic performance. However, our results revealed that Openness to experience played a significant role in differentiating paired students' academic performance. Our results also indicated that PP not only caused an increase in satisfaction and confidence levels but also brought enjoyment to the tutorial classes and enhanced students' motivation.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

INTRODUCTION

This thesis describes a series of formal experiments aimed at investigating the effects of personality traits on the effectiveness of Pair Programming (PP) as a Computer Science/Software Engineering (CS/SE) pedagogical tool. The motivation for the research questions investigated in this research is detailed in a systematic review of existing literature on PP research applied to Higher Education. This chapter introduces the thesis' background by outlining the problem statement, the research motivation, the research contributions, and the significance of the formal experiments carried out followed by a description of the overall structure of this thesis.

1.1 Background

Pair Programming (PP) has been recognized as one of the key practices in the Extreme Programming development methodology (Beck, 1999). It has become more prevalent in industry as well as in educational settings (Hanks, Wellington, Reichlmayr, & Coupal, 2008; Livermore, 2006). It involves teams of two people developing software where one acts as a *driver*, and the other as an *observer* (Williams, Kessler, Cunningham, & Jeffries, 2000). The driver is responsible for designing, typing the code, and controlling the shared resources (e.g. computer, mouse, keyboard). The *observer* or *navigator* has the responsibility for observing how the driver works in order to detect errors and offer ideas in solving a problem. Throughout their work, pairs typically alternate their roles after a certain duration (Williams & Kessler, 2002).

PP's popularity has drawn the attention of many researchers, thus causing an increase in the number of studies conducted in both industrial as well as in educational contexts (Ally, Darroch, & Toleman, 2005). A survey of organizations from a software process improvement user group showed that 72% of the organizations, from a variety of industries, have implemented the PP practice (Livermore, 2006).

Since the advent of PP many educators have trialed and endorsed its use in educational settings, most often in courses focused on learning to program or improving programming skills (Williams, Kessler et al., 2000; McDowell, Werner, Bullock, & Fernald, 2002). Early research on the use of PP as a pedagogical tool focused mainly on its ability to benefit students in terms of productivity and quality of work produced (Williams & Kessler, 2000). For example, evidence suggests that PP could enhance enjoyment (McDowell, Werner et al., 2003; Mendes et al., 2005); increase students' confidence level (Berenson et al., 2004; McDowell, Werner et al., 2003); reduce workload (Chaparro, 2005); improve course

completion rates (Nagappan et al., 2003); and facilitate students in working more efficiently on programming tasks (Cliburn, 2003; Werner et al., 2004).

In 2000, Cockburn and Williams investigated the costs and benefits of PP based on empirical evidence reported by Williams, Kessler et al. (2000), Williams (2000), and Nosek (1998). They concluded that, with an increase of only 15% in the cost of development time, PP offers significant benefits such as improving design quality (fewer defects), team communication and rapid solutions to problems, enhancing the learning process, and increasing the enjoyment of learning. They suggest that PP is a promising approach to use as a pedagogical tool due to its capability to increase learning capacity (Cockburn & Williams, 2001).

Dyba, Arisholm, Sjoberg, Hannay, & Shull (2007) conducted a systematic literature review (SLR) investigating whether existing empirical evidence would support the claims of PP being more advantageous than solo programming. They reviewed 15 studies comparing solo and PP, and involving both students and software practitioners as subjects. The general aspects investigated were related to PP's effectiveness, including "duration" (time spent to produce system), "effort" (person-hour spent), and "quality of the final product". Their meta-analysis suggested PP to be more effective than solo programming when quality and the duration to complete the tasks were the concern, but PP overall required more effort (i.e more person-hours). They also reported that it was likely that participants' expertise and task complexity might have affected the accuracy of their findings.

As PP inherently involves a social interaction between two people, investigating compatibility aspects is, in our view, very important. Previous studies reported that students who experience PP with an incompatible partner disliked the collaborative work (Layman, 2006; Thomas, Ratcliffe, & Robertson, 2003). For example, Muller and Padberg (2004) show that the performance of a pair is correlated with how comfortable the pairs feel during a pair session ("feel-good" factor). Chaparro, Yuksel, Romero, and Bryant (2005) suggest that the potential to effectively use PP is highly concerned with how compatible the students work as pairs. Such compatibility can be obtained, for instance, by matching the pair based on the similarity of students' skill levels (Chaparro et al., 2005). A pair's compatibility may also be affected by other factors such as gender, self-esteem and personality as reported by Katira, Williams, Wiebe, Miller, Balik, & Gehringer (2004).

Since students' performance may be largely affected by a pair's compatibility, it seemed relevant and applicable to examine underlying factors that may contribute to a successful pairing formation or factors that have the potential to affect the effectiveness of students' pairing. In order to realize how PP can significantly contribute as an effective pedagogical tool, a proper investigation of its implementation needs to be carried out. Our goal is not only to contribute to the body of knowledge on PP but also to improve the use of PP as an effective pedagogical tool.

1.2 Research Motivation

Over more than a decade researchers have investigated PP and its usefulness and effectiveness in both academic and industry settings (Hannay, Dyba, Arisholm, & Sjoberg, 2009; Salleh, Mendes, & Grundy, 2010). In an academic context, studies reported that PP benefits students' learning outcomes in a number of ways (as detailed in the Background section above). These benefits however do not come without a cost. The two major issues frequently highlighted in the PP literature that hinder its effective implementation in higher education contexts are scheduling conflicts (DeClue, 2003; Howard, 2006); and partner incompatibility (Ho, 2004; Layman, 2006). Such incompatibility issues might be related to psychosocial aspects such as personality and gender differences (VanDeGrift, 2004; Choi 2004). Finding a compatible or matching partner is a challenge and considered a complex issue not only in academic but also in industry settings (Sfetsos, Stamelos, Angelis, & Deligiannis, 2009).

A recent survey by Microsoft researchers has identified "personality conflicts" as the third major problem with PP, as perceived by the developers (Begel & Nagappan, 2008). Since PP is a practice involving social interaction between two people working closely together to solve programming and/or design problems, one can argue that its effectiveness can be potentially affected by human-related factors such as personality (Hannay et al., 2010; Sfetsos et al., 2006). Existing literature in Agile methods also suggests that developers' personality is one of PP's most critical success factors (Cockburn, 2002; Highsmith, 2002). Weinberg (1971) noted that "*Because of the complex nature of the programming task, the programmer's personality – his individuality and identity – are far more important factors in his success than is usually recognized*" (p. 158). Understanding how personality affects or relates to PP's effectiveness is therefore an important aspect, which has motivated us to carry out the research described in this thesis.

In relation to this, the issue of personality in PP has been addressed in a number of studies (e.g. Williams et al., 2006; Choi et al. 2008; Sfetsos et al., 2006; Hannay, Arisholm, Engvik, & Sjoberg, 2010) where their central theme was to investigate the impact of personality on the performance of teams and individuals practicing PP. Based on the results from our SLR of PP in higher education (see Chapter 2), we found evidence that only 23% of the included studies had empirically investigated factors that may affect PP's success, one of them including personality (Salleh et al., 2010). Empirical evidence from our SLR suggested that personality was one of the most common factors investigated in previous PP studies. Nonetheless, the results from these studies were inconsistent in terms of the effect or influence of personality towards pairing effectiveness (Salleh et al., 2010). This could be due to the differential set of instruments and personality frameworks used to measure personality, and the variation in the studies' context, thus making it difficult to generalize the results.

The motivation of our research was also driven by the fact that many existing PP studies employed the Myers-Briggs Type Indicator (MBTI) as a personality measurement (Salleh et al., 2010). The issues of using this personality test are highlighted in the following section and

also detailed in Chapter 3. Although MBTI was found to be very popular and widely used by researchers in the computing and business domains, there has been a rapid emerging consensus by personality psychologists on the value of the Five-Factor Model (FFM) or "big-five" as a parsimonious and comprehensive framework of personality traits (Digman, 1990; Costa & McCrae, 1992b; Burch & Anderson, 2008). Such a growing acceptance of the FFM has motivated us to employ this framework in our research. The FFM consists of five personality dimensions known as *Extraversion*, *Agreeableness*, *Conscientiousness*, *Neuroticism*, and *Openness to experience* (Costa & McCrae, 1992a).

1.3 Problem Statement

In assessing personality, the MBTI has been used as personality measure in most existing PP studies in academic settings (Salleh et al., 2010). Others have used the Keirsey Temperament Sorter (KTS) (Sfetsos et al., 2009) and most recently some studies have applied the big five or Five-Factor personality model (Hannay et al., 2010; Salleh et al., 2009; Salleh, Mendes et al., 2010a; Salleh, Mendes et al., 2010b).

The MBTI is one of the most widely-used personality assessment tools used to measure an individual's personality based on four basic dichotomous dimensions: *Extroversion (E)* vs. *Introversion (I)*, *Sensing (S)* vs. *Intuition (N)*, *Thinking (T)* vs. *Feeling (F)*, and *Judging (J)* vs. *Perceiving (P)* (Myers-Briggs et al., 1998). Thus, an individual personality type can be determined based on combination of any four of these preferences (e.g. ENFJ). However it has been argued that the MBTI is designed to measure the type of personality preferred by an individual and hence does not represent the personality traits or personality attributes of a person (Furnham, 1996; McCrae & Costa, 1989).

In the area of training and consultancy, the MBTI is commonly used as an instrument to measure personality (Furnham, 1996). The MBTI has also been widely-used by researchers in the Information Systems and Software Engineering domains (Gorla & Lam, 2004; Bradley & Hebert, 1997; Cunha & Greathead, 2007). In spite of MBTI's popularity, this instrument has been widely criticized in regard to its reliability and validity as a measurement test (e.g. Hicks, 1984; Davito, 1985; Schriesheim, Hinkin, & Podsakoff, 1991; McCrae & Costa, 1989). The MBTI instrument, which uses as its basis the psychodynamic type theory of Jungian concept, is criticized as having a number of psychometric limitations including its construct validity, and test-retest reliability which can cause bias in the interpretation of the results (Boyle, 1995; Bjork & Druckman, 1991).

The findings reported in five PP studies that investigated personality using the MBTI were quite diverse (Salleh et al., 2010), and only one study reported that pairing worked effectively for pairs of different personality types (Choi, et al., 2008). Another study by Sfetsos et al. (2006), which applied the KTS, also suggested that pairs consisting of heterogeneous personalities performed better than pairs with the same personality type. Other studies however, reported either mixed findings or found no significant effects of personality on PP (Katira et al., 2004; Katira et al., 2005; Layman, 2006; Williams et al., 2006).

Due to the inconsistencies from this evidence, it is unclear whether personality indeed has a significant effect on performance for those practicing PP. This, together with the lack of psychometric soundness of the MBTI, has led us to carry out an additional investigation on the aspect of personality's effect on PP employing the FFM. The FFM is chosen in this research because evidence shows that this personality framework is well accepted, widely assessed and extensively used by personality psychologists as well as academic personality researchers (Furnham, 1996; Conard, 2006; Burch & Anderson, 2008). The FFM adequately represents major differences between individuals and is generally considered as the most useful taxonomy for classifying personality scales (Costa & McCrae, 1992b; Barrick, Mount, & Judge, 2001).

1.4 Research Question

Based on the problem statement and research motivation aforementioned, the primary research question for this research is the following:

> "*Do personality traits affect academic performance of undergraduate students practicing pair programming*?"

We focused our investigation on the three major traits or factors from the FFM: *Conscientiousness*, *Neuroticism*, and *Openness to experience*. These factors were chosen because they were considered as the most relevant for influencing academic success in tertiary education (De Raad & Schouwenburg, 1996). Conscientiousness relates to one's achievement orientation where highly conscientious individuals are described as being diligent, hardworking, and organized (Driskell, Goodwin, Salas, & O'Shea, 2006). The level of Neuroticism determines one's ability to remain calm and composed. People who are emotionally stable (i.e. low Neuroticism) are better able to cope with stress and anxiety. Openness to experience is linked to one's imagination, aesthetic sensitivity, intellectual curiosity, and originality (Costa & McCrae, 1992a).

Our approach to answer the primary research question was to conduct a series of formal experiments in order to investigate the effects of each individual personality factor towards PP's effectiveness. These experiments involved CS undergraduate level courses at the University of Auckland. An experimental method was chosen as it provides a design that can be used for testing causal relationships (Creswell, 2003). The overview of our research process is detailed in Chapter 4, where we describe the methodology, instruments, and strategy for analyzing the data. In addition to answering the above key research question, this research also aimed to investigate a secondary research question: "*Would personality trait composition in PP affect the level of satisfaction and confidence of students when pairing?*"

1.5 Research Contributions

This research has made several contributions to the body of knowledge in the domains of Software Engineering (SE) and Computer Science (CS) education. In particular, the research contributed to the PP and SE body of knowledge by providing empirical evidence on the

effects of personality traits (from the FFM's perspective) as determinant of PP's effectiveness. The key research contributions can be described as follows:

- As part of this research, we conducted a comprehensive literature review to understand the current state of existing research on PP and most importantly to discover the gaps in the empirical knowledge of PP research conducted within higher education settings (Salleh et al., 2010). The method used to carry out the review is known as a Systematic Literature Review (SLR). The outcomes of the SLR were to benefit researchers and educators in better understanding how effective PP has been when applied as a pedagogical tool in CS/SE education system. The results were also used to further identify future PP research avenues aimed at improving PP's effectiveness or usefulness in terms of increasing students' learning. As a result, evidence from the SLR helps educators who are considering incorporating PP into a CS/SE curriculum, and the SLR's implications provide directions to researchers to conduct further studies. The SLR also contributed to support the evidence-based paradigm in SE (Dyba, Kitchenham & Jorgensen, 2005).
- In addition to the review of PP research, we also undertook a review of studies relating to personality-psychology from both educational and industrial organizational perspectives to obtain evidence in terms of the application of the Five-Factor personality model in improving team effectiveness as well as academic performance. The review delineated currently accepted personality frameworks by personality psychologists and academic personality researchers and outlines the results obtained from relevant individual studies. These knowledge areas are essential for supporting the development of our primary research question and subsequent research hypothesis.
- Further contributions of our research comprise the results from the series of formal experiments conducted to investigate the effects of personality traits on paired students' academic performance based on the FFM. These experiments, which focused on the three important personality traits from educational viewpoints, present empirical evidence and discuss the implication of studies to CS/SE educators and researchers (Salleh, Mendes, Grundy & Giles, 2009; Salleh, Mendes et al., 2010a; Salleh, Mendes et al., 2010b).
- The instruments created and used as part of this research can be employed for future replication studies. In particular, we have extended an existing tool - PALLOC software package for the purpose of facilitating the pairing allocation process. It randomly assigns students into pairs based on their levels of personality traits (low/medium/high) and generates a pairing list that can be used by the tutor. In addition, the PP survey can be used to measure students' satisfaction and confidence level when pairing.

1.6 Research Significance

The research described in this thesis was carried out using a quantitative inquiry by executing a series of formal experiments at The University of Auckland. At a high level, the significance of this research is twofold: i) it contributes to the PP and SE body of knowledge by providing empirical evidence regarding the effects of personality traits towards paired students' academic performance. This increases our understanding of the potential effects of personality, from the perspective of the Five-Factor personality model, towards PP's effectiveness as a pedagogical tool in CS/SE education; and ii) it provides evidence that can be used to ameliorate CS/SE learning in higher education institutions.

In this research, we hypothesized that the personality traits Conscientiousness, Neuroticism, and Openness to experience would affect the academic performance of paired students in CS undergraduate courses/tasks. However, some of the findings from our experiments (Conscientiousness and Neuroticism) did not support those hypotheses. Given that these findings presented a low statistical power, we argue that it would be premature to generalize such findings to a wider CS/SE population, and to conclude that the real effects of Conscientiousness or Neuroticism are indeed absent in the target population. In other words, future studies may obtain different results than ours. In our last experiment, we obtained evidence that shows a significant effect of Openness to experience on students' performance, a promising result given it also presented an acceptable statistical power level. Therefore, our results suggest that of the three personality traits investigated in this research, paired students' academic performance seems to be significantly affected by students' level of Openness to experience. Regardless of the variation in students' personality disposition, we found that students were satisfied with the pairing experience. PP also helps increased their confidence level, and motivation to learn programming and software design subjects.

The results of this research can therefore be used to better inform teachers about the implications of team personalities on academic performance when employing PP, such that their team formation approaches are influenced accordingly. The selection of personality traits as variables would also provide an advantage in overcoming the problems of pair incompatibility reported in some PP studies (e.g. Layman, 2006; Ho, 2004). Finally, we also believe that this research would be a useful addition to guide future research in PP team composition based on personality traits.

1.7 Thesis Organisation

The remainder of this thesis is organized according to the following chapters:

- **Chapter 2** begins by introducing the systematic literature review (SLR) method carried out as a foundation of this research, and then the results obtained from the review are synthesized. The SLR's findings are also discussed, taking together their implication for research and practice. Finally, the gaps in the existing body of knowledge are highlighted.

- **Chapter 3** presents a review of major personality theories based on the literature reported in the personality and psychology domains. Besides providing the motivation for selection of the personality framework applied in this research, this chapter discusses the role of personality disposition in affecting students' academic performance in tertiary institutions and team performance from the perspective of FFM.
- **Chapter 4** outlines the research methodology used in this research by describing the phases involved in conducting the experiments, the research objectives and the formulation of the hypothesis followed by the research design. Finally, a set of instruments and strategy for data analysis are also presented.
- **Chapter 5** presents the pilot experiment conducted as part of this research and the results obtained.
- **Chapter 6** describes the first experiment carried out in the 2009 Summer School, investigating the effects of the personality trait Conscientiousness on PP's effectiveness.
- **Chapter 7** describes the second experiment carried out in the first semester of 2009, also investigating the personality trait Conscientiousness.
- **Chapter 8** describes the third experiment also carried out in the first semester of 2009, and also investigating the personality trait Conscientiousness, but the experiment involved a more advanced computing course.
- **Chapter 9** describes the fourth experiment carried out in the second semester of 2009, investigating the effects of the personality trait Neuroticism on PP's effectiveness.
- **Chapter 10** describes the fifth experiment carried out in the first semester of 2010, investigating the effects of the personality trait Openness to experience on PP's effectiveness.
- **Chapter 11** analyzes the evidence gathered from each experiment, and includes the discussion on the threats to the validity of the findings, and the implications of our research for researchers and educational institutes.
- **Chapter 12** concludes the thesis by highlighting the research contributions and summarizing key directions for future work.

Chapter 2

A SYSTEMATIC LITERATURE REVIEW OF PAIR PROGRAMMING

This chapter presents the background work that informed the research detailed in this thesis. It includes the description of the methods used to carry out the literature review, and presents the detailed results of the review along with a discussion of key findings, threats to the validity of the results, and implications for future research. This chapter also includes lessons we have learnt and recommendations based on our experience in conducting the review. In order to distinguish this review from other reviews and meta-analyses reported in the literature, we provide a set of comparisons to delineate the specific contribution of this review. A subset of pair programming (PP) studies conducted in academic settings that have recently been published is summarized for further reference. Finally, a summary of this review and motivation for further empirical research conclude the chapter.

2.1 The Systematic Literature Review

This Section provides the description of the methods used to conduct the literature review detailed herein. We applied a systematic literature review (SLR) method to methodically identify, evaluate, and analyse all available evidence on pair programming (PP) research applied to the higher education context, which is the focus of this research. A SLR is defined as a process of identifying, assessing, and interpreting all available research evidence with the purpose to provide answers to specific research questions (Kitchenham & Charters, 2007). Petticrew and Roberts (2006) highlight that a SLR is a tool that aims to produce a scientific summary of the evidence in a particular area, in contrast to a “traditional” narrative review. Recent literature has shown an increase in use of the systematic review method due to the sometimes poor quality of traditional literature reviews (Dyba & Dingsoyr, 2008b). The main motivation and rationale for using a SLR is due to its comprehensive nature in performing a review, involving a rigorous method of collecting all related studies pertaining to a research question and subsequently assessing those studies in an unbiased manner.

We adopted the procedures for performing SLRs described by Kitchenham and Charters (2007). The SLR followed the procedures outlined below, some of them with iteration:

1. Formulate the review’s research question.
2. Develop the review’s protocol.
3. Identify the relevant literature by conducting a comprehensive and exhaustive search.
4. Selection of primary studies based on the inclusion/exclusion criteria.
5. Extraction of data together with studies’ quality assessment.

6. Synthesis of evidence.
7. Write up of the SLR report.

Figure 2.1 shows an overview of the SLR process, starting with the planning stage, and followed by execution of the review. During the planning stage, the SLR process begins with the preparation of a protocol to be used as a framework for the SLR. The protocol, part of the planning stage, describes the strategies to be executed when the review is carried out (Kitchenham & Charters, 2007). It specifies the research question(s), the strategy to be used for searching and storing the literature, the studies' inclusion and exclusion criteria, the strategy for assessing a study's quality, the detailed item/data to be extracted from each study, and the strategy to synthesize the evidence. In addition, the protocol developed is endorsed or approved by the main research supervisor before execution. Since the review itself is an iterative process, any changes that occurred during the actual review are updated in the protocol. The protocol for performing a SLR of empirical PP studies in higher education has been developed prior to performing the review (see Appendix A.1).

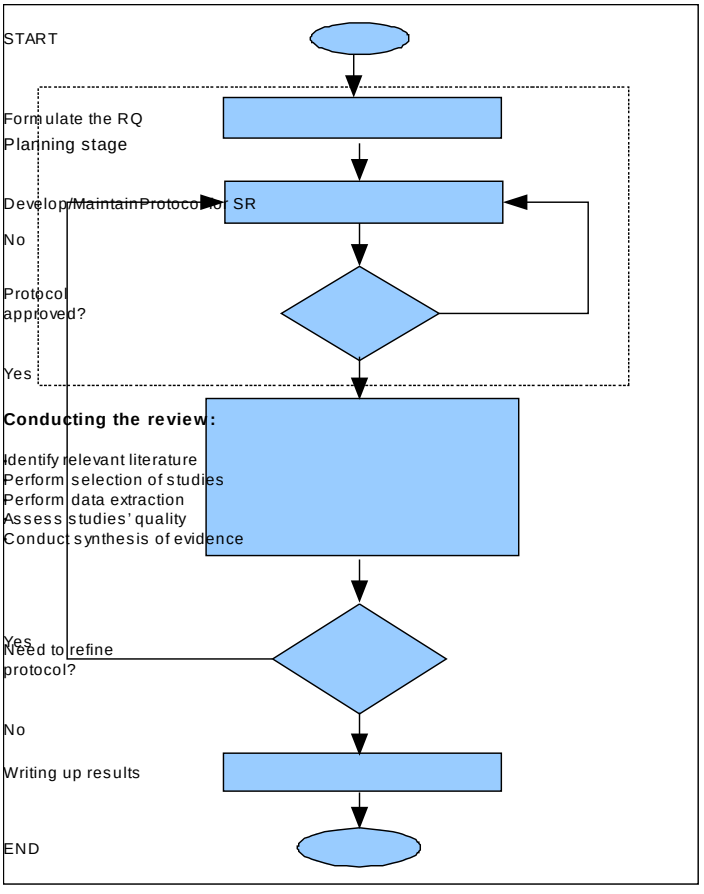


Figure 2.1 Overview of systematic review process

We carried out the SLR once the protocol was approved by the main research supervisor. The actual review begins by identifying the relevant literature required to answer the SLR's research questions. This involves searching of evidence using the search strategy detailed in the protocol. Once the relevant literature is gathered, the selection of studies takes place by

applying the inclusion and exclusion criteria defined in the protocol. Once studies have been selected and filtered based on the stated criteria, their data needs to be extracted. While extracting the data, the study quality is also assessed. The study quality assessment provides the means to appraise the strength of the evidence provided by each of the selected studies based on their methodological rigour and lack of bias (Kitchenham & Charters, 2007). The following sub-sections discuss the derivation of the research questions investigated during the review process, followed by a detailed explanation of the execution of the SLR.

2.1.1 SLR's Research Questions

In any SLR, formulating the research questions (RQ) is one of the most critical parts of the review since it drives the reasons for commissioning a review (Kitchenham, Dyba & Jorgensen, 2005). The formulation of RQs is composed of five elements known as PICOC: *Population, Intervention, Comparison, Outcomes,* and *Context* (Petticrew & Roberts, 2006). According to the definition, *population* defines the target group for the investigation (e.g. people, software). The *intervention* specifies the investigation aspects or issues of interest to the researcher(s). *Comparison* refers to the aspect of the investigation with which the intervention is being compared to. *Outcomes* define the effect of the intervention, and the *Context* describes the setting or environment of the investigation (Petticrew & Roberts, 2006). Table 2.1 shows the PICOC structure used for the SLR detailed herein. In this SLR, the intervention refers to PP studies in higher education settings, without any specific comparison treatment. This means that we have included all empirical studies that investigated PP within the Computer Science/Software Engineering (CS/SE) higher education context, regardless of whether or not they compare PP to non-PP practice.

Table 2.1 Summary of PICOC

Population	CS/SE students in higher education
Intervention	Pair programming
Comparison	None
Outcomes	PP's effectiveness
Context	Review(s) of any empirical studies of pair programming within the domain of CS/SE in higher education. No restrictions on the type of empirical study (e.g. case study) apply.

The primary focus of this SLR is to understand and identify the factors that influence the effectiveness of the PP practice for CS/SE education within a higher education environment. This includes factors affecting the compatibility of students when working in pairs. While the primary reason for using PP in industry (Cockburn & Williams, 2001; Dyba et al., 2007) is to gain benefits in terms of economic advantage (i.e. time to market, development effort, quality etc.), the type of outcomes that can benefit students' learning is what motivates educators (McDowell, Werner et al., 2003). As such, the measurement of PP's effectiveness is organized into four broad categories: academic performance, technical productivity, program/design

quality, and satisfaction (McDowell, Werner et al., 2003). Therefore, the SLR aimed to answer the following primary RQ:

Primary Question: What evidence is there of PP studies conducted in higher education settings that investigated PP's effectiveness and/or pair compatibility in CS/SE education?

In addition to this primary research question, the SLR also aimed to answer the following secondary sub-questions:

Sub-Question 1: What evidence is there regarding pair compatibility factors that affect pair compatibility and/or PP's effectiveness as a CS/SE pedagogical tool and which pairing configurations are considered as most effective?

Sub-Question 2: How was PP's effectiveness measured in PP studies and how effective has PP been when used within higher education settings?

Sub-Question 3: How was quality measured in the PP studies that used software quality as a measure of PP's effectiveness?[1]

The answers to the above RQs are presented in this chapter based on the synthesis of the evidence included in the review. The review's results detail the state of existing research on PP's effectiveness as a pedagogical tool for CS/SE teaching. The results would be beneficial in the sense that they can better-inform educators wanting to incorporate PP into a CS/SE curriculum.

2.1.2 Identification of Relevant Literature

The process of identifying the relevant literature involves a comprehensive search to be included in the review, using as its basis suitable search strings derived to ensure a wide coverage of potential sources. The strategy used to construct the search strings for this SLR was as follows (Kitchenham & Charters, 2007):

- Derive major terms used in the review questions based on the population, intervention, outcome, and context;
- List the keywords mentioned in the articles (primary studies) we already knew about (see Table 2.2);
- Search for synonyms and alternative words (see Table 2.3). We also consulted a subject librarian to seek further advice on the proper use of the terms;
- Use the Boolean OR to incorporate alternative spellings and synonyms (see Table 2.4);
- Use the Boolean AND to link the major terms from population, intervention, and outcome (see Table 2.5).

The complete search string initially used for the searching of the literature was as follows:

(student OR undergraduate) AND (pair programming OR pair-programming) AND (experiment OR measurement OR evaluation OR assessment) AND (effective OR efficient OR successful)

[1] The choice to focus on quality was due to the fact that most of the studies we already knew about measured PP's effectiveness using quality metrics.

Table 2.2 Terms derived from keywords found in PP studies

Author(s)	Year	Keywords	Index Terms/General Terms
Mendes et al.	2006	Pair programming, collaboration, software design	Experimentation, measurement
Hanks	2006	Pair programming, student attitudes, student confidence, instructor influence, empirical software engineering, computer science education	G. Terms: Experimentation, Measurement
Muller	2006	Pair Programming, preliminary studies, post-development test-cases	-
Katira et al.	2005	Pair programming, compatibility, programming teams	Management, Human Factors
Muller	2005	Pair programming, peer reviews, empirical software engineering, controlled experiment	-
Mendes et al.	2005	CS2, pair programming, collaboration, software design	Experimentation, measurement
Werner, Hanks & McDowell	2004	Pair Programming, collaboration, gender	Experimentation, Human Factors
Nagappan et al.	2003	Pair programming, collaborative environment, Computer Science education	-
Thomas et al.	2003	Pair programming, self-confidence, first year programming, CS1, closed Labs	Human Factors
Williams, Yang et al.	2002	Pair programming, collaborative learning, Computer Science education, Extreme Programming, XP	-
Cockburn & Williams	2001	Pair programming, collaborative programming, extreme programming, code reviews, people factors	-

Table 2.3 Terms derived based on synonym words

Basic terms	Alternative terms
Student	Undergraduate
Pair programming	Pair-programming (some papers use hyphen)
experiment	Measurement, Evaluation, assessment
Effectiveness	Efficient, successful

Table 2.4 Concatenation of alternative words using Boolean OR

No.	Results
1	(Student OR undergraduate)
2	(Pair programming OR Pair-programming)
3	(Experiment OR Measurement OR evaluation OR assessment)
4	(Effectiveness OR efficient OR successful)

Table 2.5 Concatenation of all possible words using Boolean AND

Results
(student OR undergraduate) AND (pair programming OR pair-programming) AND AND (Experiment OR Measurement OR evaluation OR assessment) AND (Effectiveness OR efficient OR successful)

Petticrew and Robert (2006) highlight that the two major issues in conducting a SLR search are the sensitivity and specificity of the search. The sensitivity refers to a search that retrieves a high number of relevant studies. Specificity, in turn, causes the search to retrieve a minimum number of irrelevant studies. In the preliminary search, we retrieved a very small number of articles when using the complete search as shown above (see Table 2.5). For instance, IEEEXplore, Inspec, and ProQuest each retrieved only five, three, and four articles respectively. Therefore we sought the opinion of a subject librarian regarding the appropriate use of the search string, and her advice was that a much simpler string than the one defined in the protocol should be used to enable the retrieval of more results. Therefore, we used the following

keywords:

*"**pair programming**" OR "**pair-programming**"*

which resulted on a higher number of studies retrieved from various online databases.

The primary search process involved the use of 12 online databases: *ACM Digital library*, *Current Contents*, *EBSCOhost*, *IEEEXplore*, *ISI Web of Science*, *INSPEC*, *ISI Proceedings*, *ProQuest*, *Sage Full Text Collections*, *ScienceDirect*, *SpringerLink*, and *Scopus*. The selection of online databases was based on the knowledge of databases that indexed the previous PP primary studies we were aware of, and the list of available online databases subscribed by the University of Auckland's library under the "Computer Science" subject category. Khan, Kunz, Kleijnen, & Antes (2003) recommend searching multiple databases to cater for as many citations as possible. This is because limiting the search to only a few databases might cause bias in the review. Thus, despite the list of online databases mentioned above, we also searched on the *Citeseer* website using similar keywords (i.e. "pair programming" OR "pair-programming").

From the *Agile alliance* website, we looked for a set of articles under two categories: "pair programming" and "Extreme programming". In addition, an online *Google scholar* search was used to search for full text articles. Our experiences in literature search support the suggestion by Kitchenham & Charters (2007) that it is important for software engineering researchers to identify and establish a list of relevant online databases in order to facilitate the search process.

Upon completion of the primary search phase, the identification of relevant literature continued with the secondary search phase. During the secondary search phase, all the references in the papers identified from the primary sources were reviewed. If a paper was found to be suitable, it was added to the existing list of studies qualified for the synthesis.

2.1.3 Selection of Studies

The main inclusion criteria was to only include PP empirical studies that targeted CS/SE education, and that used PP as a practice as defined by the Extreme Programming (XP) creators in 1999 (Beck, 1999). As such, the literature searching was confined to the period of 1999 to 2007. The detailed inclusion criteria comprised the following types of studies:

- Studies that investigated factors affecting the effectiveness of PP for CS/SE students.
- Studies that measured the effectiveness of PP for CS/SE students.

The main exclusion criterion comprised PP papers not targeted at CS/SE education. In addition, the following exclusion criteria were also applied during the selection of studies:

- Papers presenting unsubstantiated claims by the author(s) with no supporting evidence.
- Papers about Agile/XP describing development practices other than PP, such as test-first programming, refactoring etc.
- Papers that only described tools (i.e. software or hardware) that could support PP.

- Papers involving students outside of higher education.
- Papers that solely investigated distributed PP.
- Papers not written in English.

During the selection of studies, the title and abstract of the article were referred to in order to see whether the study complies with the SLR's inclusion and exclusion criteria. If the abstract did not provide sufficient information to decide whether the article is relevant for the SLR, the full text was then referred to.

2.1.4 Data Extraction and Study Quality Assessment

Extracting the data for synthesizing the evidence involves the comprehensive reading of a full text article to collect the data needed for the synthesis of evidence (Kitchenham & Charters, 2007). To facilitate the extraction process, we designed a data extraction form (see Appendix A.2). The data extraction form consists of items used to answer the research questions as well as items used to measure a study's quality. The data items extracted from each study were:

- The full reference of the paper (author, paper title, year, type and source of publication).
- Context of the study (aim of study, study setting, strategy, subject, description of tasks, courses involved, and duration of study).
- The method used for pair allocation and whether the study compared pairs to individuals.
- List of variables investigated in the study (dependant, independent, and confounding variables) and how these variables are operationalized.
- Hypotheses or research question(s) used in the study.
- The type of study design used.
- Compatibility factor(s) addressed or investigated in the study.
- The criteria used to measure effectiveness (i.e. how was PP's effectiveness measured and investigated in the study).
- The criteria used to measure the quality aspects of PP.
- Method(s) of data analysis and the statistical method(s) applied in the study.
- Whether the effect size is reported or calculable (based on the reported statistics).
- Summary of findings from the study.

Assessing the study quality is an important component of a SLR which is considered as critical in order to evaluate the potential of research bias (Kitchenham & Charters, 2007; Petticrew & Robert, 2006). Petticrew and Robert (2006) describe the process of assessing a study's quality as a means to achieve "internal validity" by identifying the extent to which a study is free from major methodological biases. Each study included in a SLR should be critically appraised to determine its appropriateness in answering the research questions. These include assessing the study's methodology or design, and methods to analyze the results (Khan et al., 2003). For this purpose, we reused some of the questions proposed in

the literature in order to measure the quality of both quantitative and qualitative studies (Leedy & Ormrod, 2005; Spencer, Ritchie, Lewis, & Dillon, 2003; Crombie, 1996; Fink, 2005; Greenhalgh, 2000). Based on these questions, a quality checklist comprising of seven general questions (see Table 2.6) was designed to be answered using the following ratio scale:

Yes = *1 point;* No = *0 points;* Partially = *0.5 point*

Table 2.6 Study quality checklist

Item	Answer
1. Was the article refereed? (Leedy & Ormrod, 2005)	Yes/No
2. Were the aim(s) of the study clearly stated? (Crombie, 1996; Fink, 2005)	Yes/No/Partially
3. Were the study participants or observational units adequately described? For example, students' programming experience, year of study etc. (Petticrew & Roberts, 2006; Greenhalgh, 2000)	Yes/No/Partially
4. Were the data collections carried out very well? For example, discussion of procedures used for collection, and how the study setting may have influenced the data collected (Petticrew & Roberts, 2006; Spencer et al., 2003; Fink, 2005; Greenhalgh, 2000)	Yes/ No/Partially
5. Were potential confounders adequately controlled for in the analysis? (Fink, 2005)	Yes/No/Partially
6. Were the approach to and formulation of the analysis well conveyed? For example, description of the form of the original data, rationale for choice of method/tool/package (Spencer et al., 2003; Fink, 2005; Greenhalgh, 2000)	Yes/No/Partially
7. Were the findings credible? For example, the study was methodologically explained so that we can trust the findings; findings/conclusions are resonant with other knowledge and experience (Spencer et al., 2003; Petticrew & Roberts, 2006; Greenhalgh, 2000)	Yes/No/Partially

The "*Yes*" answer is given if the stated quality criteria is explicitly mentioned or described in the study; the "*Partially*" answer is given if the stated quality criteria was mentioned implicitly or only partially defined in the study. Finally, the "*No*" answer is given if the study clearly lacked the particular quality element. Therefore, the total quality score for each study ranged between 0 (very poor) and 7 (very good). The quality assessment however did not intend to exclude any study and was used only as a quality benchmark. The exception was a study in which we found that the quality was so poor that it was unable to provide sound evidence to the SLR.

In order to validate the data extraction process, a random sample comprising 20% of the total number of primary studies was selected randomly and had their data extracted by a review team consisting of the main researcher and her primary supervisor. The extracted data was later compared in a review meeting attended by the review team. Whenever there were differences in the data extracted, which for a given paper, was never beyond 10-15% of the total data extracted, these differences were discussed until a consensus was reached. We did

not measure inter-rater agreement since the review aimed to reach an absolute consensus on the sample used (Kitchenham, Mendes & Travassos, 2007). The lesson learnt from the review meeting is that it is expected to minimize, if not remove, the bias with the data extraction for the remaining papers. The study author(s) were contacted where information was unclear.

2.2 The Results of the Review

In this section, the synthesis of evidence of the SLR is presented, beginning with the analysis of the literature search results. The breakdown of literature searches from online databases is shown in Table 2.7. During the selection process, the Scopus database was chosen as the baseline database due to its reputation as the largest abstract and citation database (Elsevier, 2008). In addition, each article retrieved from the other databases was compared with the existing list of papers accumulated from Scopus' screening process, in order to avoid duplication.

Table 2.7 Breakdown of literature searches

No	Database Name	Total record found (A)	Total # Duplicate (B)	Total # included for screening (C=A-B)	Total # Not relevant (after screening TI and ABS) D=(using C)	Total # to be reviewed (after screening TI and ABS) (E = C – D)
1	Scopus	129	-	129	50	79
2	IEEEXplore	70	44	26	15	11
3	ISI Web of Science	43	31	12	9	3
4	INSPEC	148	108	40	21	19
5	ScienceDirect	33	4	29	25	4
6	ISI Current Contents	22	21	1	-	1
7	ISI Proceedings	69	63	3	2	1
8	SpringerLink	44	36	8	6	2
9	ProQuest	53	13	40	25	15
10	EBSCOhost	45	28	17	13	4
11	ACM	58	32	26	14	12
12	Sage Full Text	9	7	2	1	1
13	Citeseer	59	39	20	19	1
14	Agile alliance	30	4	26	26	0
	TOTAL			**379**	**226**	**153**

(Note: TI = Title; ABS = Abstract)

The initial phase of the search process identified 379 empirical studies using the *"pair programming OR pair-programming"* search term. Of these, only 153 studies were potentially relevant based on the screening of titles and abstracts. Each of these studies was filtered according to the inclusion and exclusion criteria before being accepted for the synthesis of evidence. If the titles and abstracts were not sufficient to identify the relevance of a paper, the full articles were used during the selection process. We also checked if there were any duplicate studies whenever very similar studies were published in more than one venue. The inclusion of duplicate studies would inevitably bias the result of the synthesis (Khan et al., 2003). Therefore, careful assessment was carried out to detect any duplicate studies, by comparing their data, the study's outcomes, and their period.

After a detailed assessment of abstracts and full texts, and exclusion of duplicates, 73 studies from the primary search phase (48% of 153 studies) were accepted for the synthesis

of evidence (see Figure 2.2). The secondary search phase further identified another five studies; however, after their detailed assessment, only one was found relevant for the SLR. Therefore, in total, data from 74 studies were extracted for the synthesis of evidence (see Appendix A.3 for the list of included studies).

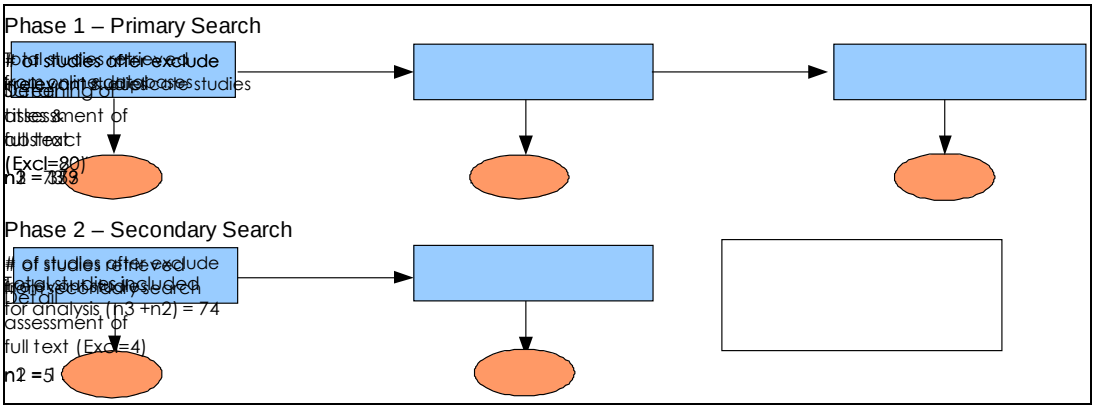


Figure 2.2 Identifying relevant literature

In order to validate the coverage of the search process, we had manually compared the list of articles retrieved from the search with the list of primary studies we already knew about. Table 2.8 provides the coverage of the search process, which suggests that the search coverage was highly reliable. An article written by Nosek (1998) was the only paper that did not appear during the identification of relevant studies. The reason was that the keyword “pair programming” or “pair-programming” did not exist in the paper. Instead, the author of the article used the term “collaborative programming” to indicate a team consisting of two programmers working collaboratively. The paper had also been found during the secondary search phase, but rejected for the analysis since it did not satisfy the inclusion criteria.

Based on the research classification by Wohlin et al. (2000) and Creswell (2003), an analysis of the type of research approach used in these studies is shown in Figure 2.3. Formal experiments were found to be the most popular research approach used (59%), when compared with other approaches such as surveys, case studies, mixed-methods, and qualitative studies.

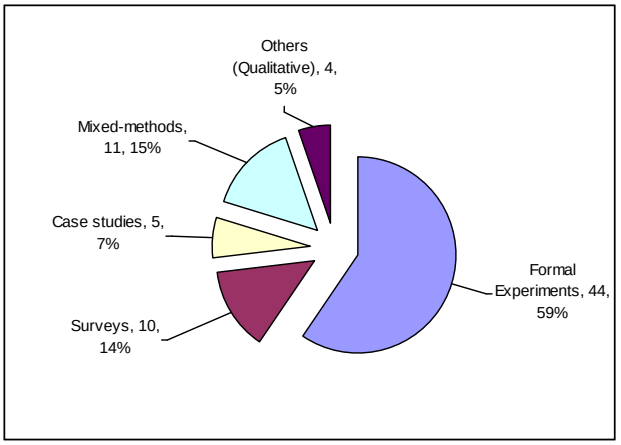


Figure 2.3 Studies by research approach

Table 2.8 Coverage of search process

			Scopus	IEEE	Web of Science	INSPEC	Science Direct	Current Content	ISI Proc.	Springer	Pro Quest	EBSCO host	ACM	Sage
Number of papers retrieved			121	70	42	148	33	22	69	44	52	45	58	9
Author	Year	Expected to see in:	Did search find this paper? (Yes/No)											
Nosek	1998	ACM											NO	
Williams & Kessler	2000	IEEE	YES	YES	YES	YES					YES	YES	YES	YES
Cliburn	2003	ACM											YES	
DeClue	2003	ACM											YES	
McDowell, Werner et al.	2003	ACM	YES	YES		YES							YES	
Nagappan, Williams, Ferzli et al.	2003	ACM	YES			YES							YES	
Williams, McDowell et al.	2003	IEEE		YES		YES			YES				YES	
Hanks et al.	2004	ACM	YES			YES							YES	
Katira et al.	2004	ACM	YES			YES							YES	
Mendes et al.	2005	ACM	YES			YES							YES	
Howard	2006	EBSCOhost				YES						YES		
Mendes et al.	2006	ACM	YES			YES							YES	
Williams et al.	2006	IEEE	YES	YES		YES			YES				YES	

The primary studies included in the SLR were also categorized based on the year of publication and research approach. As can be seen in Table 2.9, the number of studies increased yearly since 2002 until 2006; however from 2006 onwards they have decreased.

Table 2.9 Studies of pair programming by research method and year

Year/Study type	2000	2001	2002	2003	2004	2005	2006	2007
Formal Experiments	0	1	2	7	5	12	11	6
Surveys	1	0	1	3	1	1	3	0
Case studies	0	0	1	0	1	1	0	2
Mixed-methods	1	0	1	1	3	1	2	2
Others (Qualitative)	0	0	0	0	1	3	0	0
TOTAL	**2**	**1**	**5**	**11**	**11**	**18**	**16**	**10**

Table 2.10 presents the quality scores for all primary studies, and suggests that most studies were of a good quality threshold, with 36 studies (49%) being of good quality, and 20 studies (27%) of very good quality. There was one study that was of very poor quality, and therefore was removed from the analysis phase. It provided no details on its research methodology and, as a consequence, we could not ensure its results were reliable and useful as evidence. We have also contacted the researchers who authored that paper, but received no reply. Thus, in the end only 73 studies were included in the analysis of evidence.

Table 2.10 Quality scores

Quality Scale	Very Poor (<2)	Poor (2 – < 3)	Fair (3 – <5)	Good (5 – <= 6)	Very Good (>6)	Total
Number of studies	1	0	17	36	20	74
Percentage (%)	1%	0%	23%	49%	27%	100%

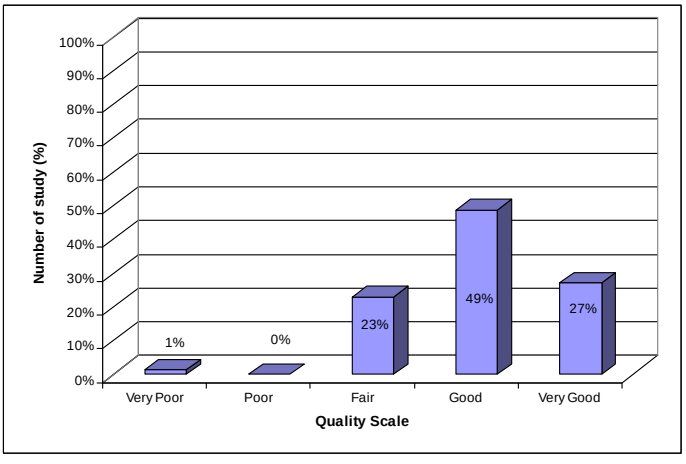


Figure 2.4 Statistics on Study Quality Scores

In the following section the results for the SLR's main research question and its three sub-questions are described. Each study is identified as S*m*, where *m* represents the study's number (see Appendix A.3 for the list of studies used in the SLR).

2.2.1 Synthesis of Evidence

The primary research question (RQ) of the SLR is ***"What evidence is there of PP studies conducted in higher education settings that investigated PP's effectiveness and/or pair compatibility for CS/SE education?"***

The SLR initially included 74 PP studies conducted in higher education settings that investigated the use of PP by undergraduate and graduate CS/SE students. Of the 74 studies, one was excluded from the analysis due to its poor quality thus leaving only 73 studies for the synthesis of evidence. The context of investigation varied via the comparison of PP to other practices, such as solo programming, side-by-side programming, peer-review inspection, and application of the practice to software design tasks (i.e. pair designing). Studies that investigated PP's effectiveness also covered other aspects of PP such as pair formation and its relation to pair compatibility.

The SLR's ultimate goal was to understand how PP affects students' learning outcomes in order to improve their academic achievement, technical productivity, program quality, and learning satisfaction. Of the 73 studies included in the analysis, 17 (23%) investigated pair compatibility as a factor believed to have a bearing on PP's effectiveness. Seventy (96%) investigated PP's effectiveness using either a quantitative or a qualitative approach, and 32 (44%) investigated quality aspects as a measure of PP's effectiveness. The following subsections detail the SLR's synthesis of evidence.

2.2.2 Sub-question 1 – Compatibility Factors

"What evidence is there regarding pair compatibility factors that affect pair compatibility and/or PP's effectiveness as a CS/SE pedagogical tool, and which pairing configurations are considered as most effective?"

Compatibility factors are factors believed to influence the compatibility of students when working in pairs. Altogether, 14 factors were investigated in 17 studies that looked into how they affected or correlated with PP's effectiveness and/or pair compatibility. Table 2.11 presents the compatibility factors, the studies that investigated each factor, and whether the factor had a positive, negative, no effect, or mixed effect. The summary of findings for both quantitative and qualitative studies, used to answer this research question, is available in Appendix A.4.

As can be seen from Table 2.11, personality type and actual skill level were the two factors most commonly investigated in PP studies. In terms of personality, the two studies with positive findings reported that paired students of different personality types performed better when compared with paired students of similar personality type [S50], [S73]. While most studies that investigated the effects of personality type did not produce significant findings, there was agreement that paired students of different personalities tended to perform better than pairs of similar personalities [S28], [S32], [S63].

Table 2.11 List of factors investigated in PP studies

No.	Factor	Total studies	Significant positive effect	Significant negative effect	No significant effect	Mixed findings
1	Personality	9	S50, S73	-	S13, S23, S29 S32 S74	S28, S63
2	Actual skill level	10	S8, S11, S15, S28, S29, S58, S63, S68, S74	*S11, *S58	S42	-
3	Perceived skill level	4	S14, S28, S29, S63	-	-	-
4	Self-esteem	3	-	-	S8, S29, S63	-
5	Gender	2	S29	-	S73	-
6	Ethnicity	1	S29	-	-	-
7	Learning style	2	-	-	S32, S63	-
8	Work ethic	2	S63	-	S32	-
9	Time management ability	1	-	-	S63	-
10	“Feelgood“ factor	2	S42, S68	-	-	-
11	Confidence Level	2	S22	S54	-	-
12	Type of role	1	-	-	S14	-
13	Type of tasks	1	S14	-	-	-
14	Communication skills	1	-	-	S73	

*S11 and S58 both reported that skill level had positive & negative effect on the PP’s effectiveness. Pairs consisting of similar skill can benefit students, but pairs consisting of very different levels seem ineffective.

Six out of the nine PP studies that investigated personality type employed the Myers-Briggs Type Indicator (MBTI) as a personality assessment method [S13], [S28], [S29], [S32], [S63], [S73]. Only one study applied NEO Personality Inventory (NEO-PI) to investigate the relationship between programmers’ personality and PP’s effectiveness [S23]. The study found that the personality of an individual programmer does not have a significant effect on PP’s effectiveness, but this may not be the case when looking at the combination of personalities in a single pair. Other than MBTI and NEO-PI, the Keirsey Temperament Sorter (KTS) and Revised Eysenck Personality Questionnaire (EPQ-R) were used in two studies to measure the personality and temperament types of pair developers [S50], [S74]. Sfetsos et al. (2006) [S50] report that pairs of mixed-personalities and temperaments achieve better scores than pairs of similar personality. On the contrary, Gevaert (2007) [S74] found no significant correlation between personality type and PP’s effectiveness.

Seven out of the ten PP studies regarded paired skill level as one of the determinant factors of PP’s effectiveness [S8],[S11],[S15],[S28], [S29],[S63],[S58]. The two categories of skill level used were actual and perceived skill. The actual skill level was determined based on programming experience, academic background, and students’ academic performance (i.e. assignments, exams, and project scores), whereas perceived skill level was measured subjectively according to the skill of a student’s partner relative to their own perceived skill (i.e. “better”, “about the same”, or “weaker”). The consensus from these seven studies was that PP worked best when the pair consists of students of similar skill level. However, two correlation studies [S42], [S68] show contradictory findings on the association between students’ skill level and PP’s effectiveness. Muller and Padberg (2004) [S42] report that there is no correlation between the two variables, but Madeyski (2006b) [S68] refutes this finding.

The two studies that investigated the effect of gender differences on pair compatibility produced contradicto ry findings [S29], [S73]: Choi (2004) [S73] reports that gender is not a significant factor to influence pair compatibility whereas Katira et al. [S29] found gender is a factor likely to determine pair compatibility. Katira et al. (2005) [S29] report that pairing students of different gender would lead to incompatible pairs and that pairing female students would very likely result in a compatible pair. Three studies [S28], [S29], [S63] that investigated the effect of self-esteem discovered that paired students' self-esteem did not influence pair compatibility.

Katira et al. (2005) [S29] investigated ethnicity as a compatibility factor by classifying students as either belonging to a majority or minority ethnic group. Their results show that students from minority ethnic groups are more likely to pair with students who are also from minority ethnic groups, but not necessarily the same group. In the study, the effects on pair compatibility when pairing students belonging to the same ethnicity were not investigated. The study also does not report the results of pair compatibility on male students and majority ethnic groups due its focus on addressing the issue of low representation of minority and female students in CS.

The two studies that investigated the effect of Felder-Silverman learning style reported that learning style did not significantly affect pair compatibility or the perception of students towards the pairing experience [S32], [S63]. In terms of work ethic, Williams et al. (2006) [S63] report that pairing students of similar work ethic enhances pair compatibility, and Layman (2006) [S32] reports that students' perception towards pairing is positive regardless of their work ethic. Williams et al. (2006) [S63] also investigated students' time management ability and found that it has no effect on pair compatibility.

In 2004, Muller and Padberg [S42] coined the term "feel-good" which refers to how comfortable pairs feel during the PP session. They report that the feel-good factor is positively correlated with a pair's performance (measured by the time spent). Madeyski (2006b) [S68] had similar findings where a positive correlation between the feel-good factor and pair performance (quality of software) was found.

Very few PP studies have investigated confidence, communication level, type of role and tasks. Thomas et al. (2003) [S54] report that performance increased when students of similar confidence were paired. Nevertheless, students who considered themselves as "code warriors" (i.e. high confidence level students) preferred to work alone and enjoyed PP less. This contradicts the findings reported by Hanks (2006) [S22], where students of higher self-confidence enjoyed PP the most. Chapparo et al. (2005) [S14] report that the type of task significantly affects the perceived effectiveness of PP. Paired students were found to have a preference for program comprehension, re-factoring, and coding more than debugging tasks. In terms of communication skills, Choi (2004) [S73] reports that these have no impact on pair compatibility.

Researchers at North Carolina State University conducted a two-phased study between 2002 and 2005 to investigate factors believed to have an influence on pair compatibility. Table

2.12 summarizes the findings. In these studies [S28], [S29], [S63], the experiments involved undergraduate and graduate CS students in three courses: Introduction to Programming (CS1), Software Engineering (SE), and Object Oriented Languages and Systems (OO). Synthesis of evidence showed some divergence in the findings from these studies. For instance, results were contradictory between CS1 and SE courses when students were paired according to different personality type, similar actual skill level, and self-esteem [S63]. The perceived skill level was the most influencing factor in determining pair compatibility. These studies however did not provide evidence stressing pair compatibility as an important criterion determining PP's effectiveness.

Table 2.12 Compatibility of student pair programmers

	Were the pairs compatible? (Yes/No)								
Courses	**CS1**			**SE**			**OO**		
Factors/Study	[S28]	[S29	[S63]	[S28]	[S29]	[S63]	[S28]	[S29]	[S63]
Personality Type	Yes	-	No	No	No	Yes*	-	No	No
Perceived Skill	Yes	-	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Actual Skill	No	-	No	No	Yes	Yes*	Yes	Yes	No
Self-esteem	No	-	No	-	-	Yes*	-	-	No
Gender	-	-	-	-	Yes	-	-	Yes	-
Ethnicity	-	-	-	-	Yes	-	-	Yes	-
Work ethic	-	-	-	-	-	Yes	-	-	-
Time Management skill	-	-	-	-	No	-	-	-	-
Learning Style	-	-	-	-	-	Yes*	-	-	-

*(Note: * Indicates partial support)*

The second part of sub-question 1 investigated the most effective method of pair formation from the viewpoint of pair compatibility and/or pair effectiveness. Table 2.13 presents the analysis of effective pairing formation. We presented our evidence based on the ranking of the number of studies with corroborating findings relating to pairing formation. The *actual skill level* was ranked highest (seven studies) [S8], [S11], [S15], [S28], [S29], [S58], [S63] followed by *perceived skill level* (four studies) [S14], [S28], [S29], [S63]. These studies report students should be paired with a partner of a similar skill in order to achieve greater pair compatibility or pair effectiveness. Personality type was ranked third, where two studies report that students should be paired with a partner of different personality [S50], [S70].

In terms of quality assessment, the average quality score obtained by the empirical studies used to answer this sub-question was 5.1, with the highest quality score being 6.5. Of 17 studies, 12 were rated as having good quality of experimental design and analysis.

Table 2.13 Summary on effective pairing formation

Compatibility factor	Study(s)	Pairing formation	Findings
Actual skill	S8	Similar educational background	The academic background of a pair's component affects the knowledge built. Coupling two different academic backgrounds does not seem to improve the performance.
	S11, S15, S29, S28, S63	Similar midterm/GPA/course grades	Pairs of similar or not very different level of competency were effective [S11]. 92.3% of students responded that PP made them work better with others. Their exam scores were higher compared with those from previous semesters (S15). S29 identified that students had a preference to pair with a student of similar actual skill (based on SAT/GRE/GPA scores).
	S58	Similar programming skills	PP works best when the programmers are of slightly different skill level, but the gap should not be too broad.
Perceived skill	S14	Paired students with matching skills	Skill level appeared to have a strong influence in the success of PP sessions. The skill level gap between the partners should not be too broad.
	S28, S29	Similar perceived skill or technical competence level.	Compatibility was highly affected by the perceived skill of a student's partner.
	S63	Similar or higher skill level.	Students preferred to pair with a partner they perceived to be of similar or higher skill level.
Personality type	S73	Different personality type.	Students seem more compatible with a partner of a different MBTI personality type [S28]. Paired students with diverse personality performed significantly better than the pairs of similar personality in terms of code productivity & code design (S73).
	S32	Paired the extravert students.	An extrovert student is highly sociable compared to an introvert student; thus extroverts favor collaborative work.
	S50	Mixed-personalities and temperaments.	Pairs of mixed-personalities and temperaments showed better performance and collaboration-viability. They achieve better points on assignments and shorter time to complete the tasks.
Gender	S29	Pairs of female students.	Pairs of female students will likely result in a compatible pair. Paired students of mixed gender reported to be less likely compatible.
Ethnicity	S29	Pairs of minority students.	The study reported that a pair with only minority students is more likely to be compatible. Paired students with the same gender were also reported to be more comfortable working with each other.
Learning Style	S32	Pair of sensors.	Students of sensors Felder-Silverman learning style prefer to work in a group setting. Students with a reflective learner scale and those who considered themselves strong coders disliked pairing.
	S63	Paired students of sensor and intuitor learning style.	Pairing a sensor and an intuitor lead to a very compatible pair.
Confidence level	S54	Paired students with similar or not very different level of confidence.	Students with a reasonably high self-confidence did not enjoy the PP session. Paired students of similar confidence level seemed to improve performance.
	S22	Paired of high confidence students	Students with higher self-confidence enjoyed PP the most compared with students with less confidence.
Work ethic	S63	Paired students with similar work ethic.	Students preferred to work with someone who had similar intention to success in the course as themselves.
Time management ability	S32	Paired students of lower time management ability.	Students with higher time management ability showed the tendency to work alone.

2.2.3 Sub-question 2 – Measure of PP's Effectiveness

"How was PP's effectiveness measured in PP studies and how effective has PP been when used within higher education settings?"

The effectiveness of the PP practice within a higher education setting was measured using various factors, organized into four categories: *technical productivity*, *program/design quality*, *academic performance*, and *satisfaction*. Technical productivity, measured by 31 (44%) of the

70 studies, was the most common method used to assess PP's effectiveness, followed by program/design quality (30 studies, 43%). A subset of 16 studies (23%) evaluated PP's effectiveness based on students' academic performance in final exams, midterms, assignments, projects, and course grades. Besides the objective measurements, PP's effectiveness was also evaluated subjectively in 23 studies (33%) using the perceived satisfaction of students experiencing the PP sessions (see Table 2.14).

Table 2.14 Categories of metrics to measure PP's effectiveness

Categories	Metrics used to measure effectiveness	Sig. +ve effect	Sig. –ve effect	No significant effect	Mixed findings
Technical productivity 31 studies (44%)	Time Spent	S7, S9, S30, S31, S33, S49, S52, S53, S60, S4, S51	S65	S19, S25, S38, S42, S44, S46, S47	-
	Knowledge & Skill transfer	S3, S6, S8, S26, S27	-	-	-
	Task performance & Code accuracy	S50	S18	S43, S74	-
	Number of solution that satisfy test cases	-	-	S23, S57	-
	Number & types of problem	-	-	-	S71
Program/ design quality 30 studies (43%)	Design/Project scores	S2, S9, S30, S39, S59	-	-	S62
	Lines of Code		-	S47, S25	S57
	OO Design Quality	S5,	-	S35, S25	S21
	Number of test cases passed/failed	S17, S33, S53 S60	-	S44 S49	-
	Expert opinion	S38, S52, S73	-	S13	S45
	Number of defects	S55	-	S47	S64
	NATP	S68	-	S34, S46	-
	Code coverage thoroughness	-	-	S36	-
	Standard Quality model	-	-	-	S1,
Academic Performance 16 studies (23%)	Assignment scores	S38, S39, S40, S41	-	-	S37
	Final exam scores	S48, S39, S40, S41	-	S19, S54	S37, S38
	Midterm scores	S48,	-	-	-
	Quiz scores	-	-	S19	-
	Project scores	S59	-	S20	-
	Test scores	S30, S40, S41	-	-	-
	Course grade	S15, S39, S40, S41	S24	S19	S61, S62
	Course completion rate	S12, S39	-	-	-
	Retention rate	-	-	-	S37
Satisfaction 23 studies (33%)	Pair formation, increased knowledge & confidence, positive attitude about collaboration, enjoyment, social interaction	S2, S10, S14, S16, S18, S19, S20, S21, S22, S31, S32, S39, S40, S41, S52, S56, S57, S59, S62, S66, S70, S72, S74	-	-	S69

Of the 31 'technical productivity' studies, 19 [S4], [S7], [S9], [S19], [S25], [S30], [S31],[S33], [S38], [S42], [S44], [S46],]S47], [S49], [S51], [S52], [S53], [S60], [S65] used "time spent" as a measure of PP's effectiveness. Of these, 11 studies [S4], [S7], [S9], [S30], [S31], [S33], [S49], [S51], [S52], [S53], [S60] report that paired students completed assigned

tasks in a shorter duration than solo students. However, 7 studies report that PP incurs additional cost or requires more effort (in person hours) because it requires two heads in solving a task [S25], [S65], [S60], [S52], [S53], [S46], [S47]. Some studies do not report the total effort as they included only the time taken to solve the task.

PP studies that measured PP's effectiveness using quality attributes (30 studies) focused on either the internal or external code quality. The lines of code and the number of test cases passed were the two most common methods employed. PP's effectiveness has also been investigated subjectively by means of students' perception towards their satisfaction when using PP. Findings showed a positive attitude towards working collaboratively with another student; however, scheduling conflicts and incompatible partners were the two major problems highlighted in the literature [S24], [S32], [S66], [S69], [S56], [S16].

Of the 45 studies that compared PP to solo programming, 31 report that PP led to an improved performance in terms of technical productivity and satisfaction. Although the findings regarding PP's effectiveness in program/design quality and academic performance varied considerably (see Table 2.15), the majority of studies (8 out of 16) report a significant positive effect of PP towards academic performance.

Table 2.15 PP's effectiveness (PP vs Solo)

Comparison	Technical Productivity	Program quality	Academic Performance	Satisfaction
1) PP is better than solo	S3, S4, S6, S7, S8, S9, S30, S31, S33, S51, S52, S53, S60	S2, S5, S9,S17, S30, S33, S38, S39, S53, S55, S60, S68	S12, S15, S30, S39, S40, S41, S48, S59	S2, S14, S16, S18, S19, S20, S21, S31, S40, S41, S52, S57, S62, S66, S70, S72, S74
2) PP is similar to solo	S19, S23, S25, S43, S46, S47, S57, S74	S25, S34, S35, S36, S44, S46, S47	S19, S20, S54	-
3) PP is worse than solo	S18, S65	-	S24	-
4) Mixed-findings	S71	S1,S21, S45, S57,S64	S37, S38, S61, S62	S69

Figure 2.5 suggests that PP was a more effective technique compared with solo programming. In terms of satisfaction almost all studies reported similar findings where paired students presented greater satisfaction and enjoyment when using PP. Of 24 studies, 13 report that paired students achieve better "technical productivity" than the solo students, whereas 8 report that there is no difference in performance. PP's effectiveness measured by program quality show a variation of findings: 12 studies report a positive effect of PP, 7 studies report that there is no difference in program quality between PP and solo students, and 5 studies report mixed findings.

The range of quality scores for the studies included in measuring PP's effectiveness varied between medium and high. The average quality score was 5.8, with 20 out of 70 studies (29%) obtained a very good score; 35 studies (50%) were of good quality, and only 15 studies (21%) obtained a fair score.

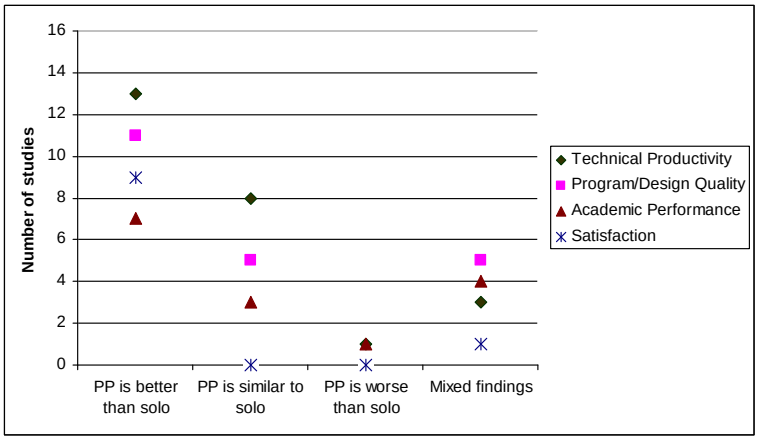


Figure 2.5 Studies' findings on PP's effectiveness

2.2.3.1 Meta-analysis on PP's Effectiveness

In an attempt to quantitatively combine the results from empirical studies using meta-analysis, Pickard, Kitchenham, and Jones (1998) suggest that only studies that have comparable measures are eligible to be included in a meta-analysis. Based on the SLR's results, we found that PP's effectiveness was measured using various types of metrics. Thus, in order to perform a meta-analysis a specific subset of measure need to be selected, such as the final exam score or the success rate of paired and solo students. The meta-analysis is then carried out to measure the effects of the intervention (PP) towards academic performance.

In the SLR, only six studies that reported their statistical results were applicable for a meta-analysis. The data reported in these six studies [S38], [S39], [S40], [S41], [S61], [S62] were used to carry out two meta-analyses: one of PP's effectiveness on final exam scores of paired and solo students (MA1), and another of PP's effectiveness on assignments' scores (MA2). MA1 showed a standardized effect size of 0.16, calculated using Hedges' *g* statistic (Rosenthal & DiMatteo, 2001). Here, the standardized mean difference under the fixed effects model was used as the measure of effect size. Effect size was calculated based on the difference between two means (final exam scores of paired and solo students) divided by the pooled standard deviation, adjusted for small sample bias (Kampenes, Dyba, Hannay, & Sjoberg, 2007). The formula to calculate the Hedges' g is defined as below (1):

$$Hedges'\,g = \frac{\overline{X}_1 - \overline{X}_2}{S_p} \qquad (1)$$

And the pooled standard deviation is calculated based on the standard deviations in both pair (s_1) and solo (s_2) groups, as defined below (2):

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}} \qquad (2)$$

The forest plot in Figure 2.6 shows the result of the first meta-analysis (MA1). The small box indicates the point estimates of effect size in a single study whereas the horizontal line that crosses each study represents the confidence interval for the study's estimate. The

diamond at the bottom of the plot represents the pooled effect or the average effect size after pooling all studies. The pooled result from this meta-analysis suggests that the effects of PP are considered small (i.e. effect size of 0.16) in terms of its practical significance, or meaningfulness in improving students' performance in final exams, compared with solo programming. In this regard, we employed the effect size category from Kampenes et al. (2007), which classifies effect size into either small (effect size of 0.000 – 0.376), medium (effect size of 0.378 – 1) or large (effect size of 1.002 – 3.40). Note that some of the studies reported their statistical results for several experiments conducted in various academic semesters, thus were treated each as a separate study in the meta-analysis (e.g. the three experiments in McDowell et al. [S38] are denoted as S38a, S38b, and S38c, respectively).

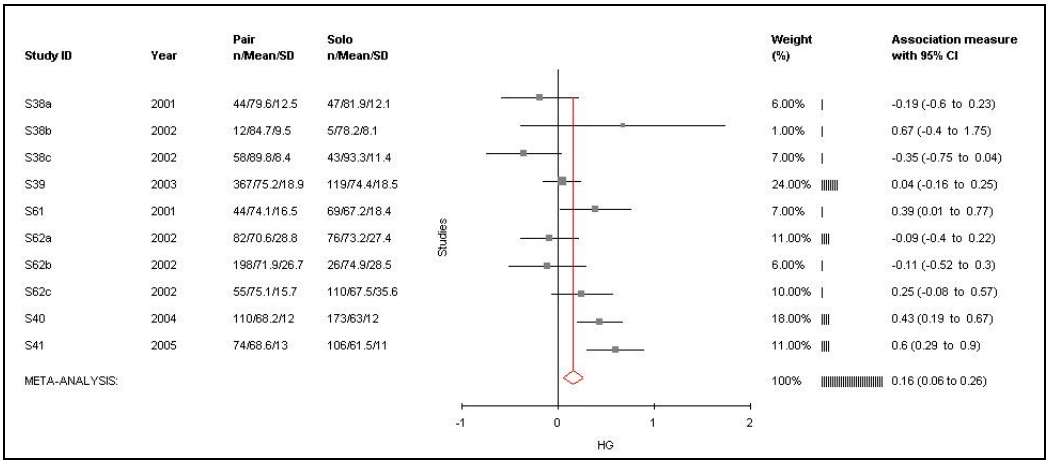


Figure 2.6 Meta-analysis of PP's effectiveness on students' final exam scores

The second meta-analysis, MA2, showed a medium effect size (see Figure 2.7). The pooled effect size of 0.67 suggests that PP was beneficial and effective in helping students getting better scores in their assignments. In both meta-analyses, the software MIX version 1.7 was used for performing both meta-analyses and generating the forest plots (Bax, Yu, Ikeda, Tsuruta, & Moons, 2006; Bax et al., 2008). The validity of MIX as a software that can perform a comprehensive meta-analysis is reported by Bax et al. (2006).

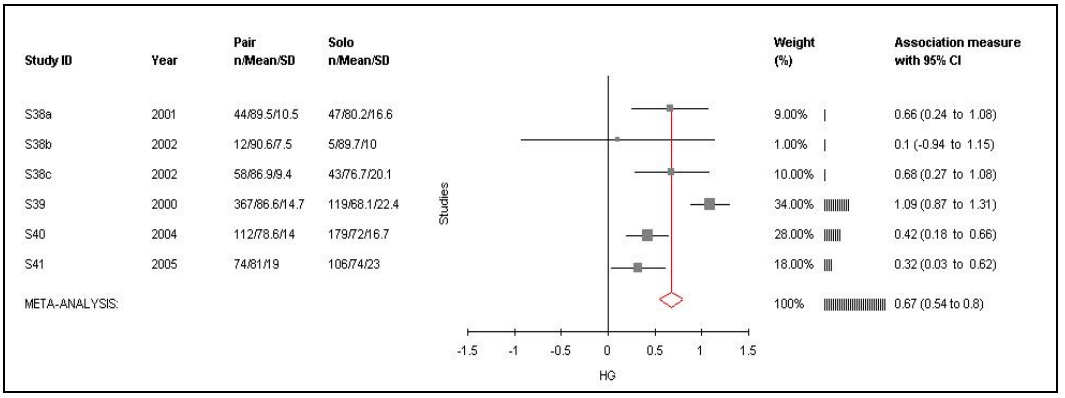


Figure 2.7 Meta-analysis of PP's effectiveness on programming assignments

2.2.4 Sub-question 3 – Measure of Quality

"How was quality measured in the PP studies that used software quality as a measure of effectiveness?"

PP is reported to benefit its users by improving software design quality (Cockburn & Williams, 2001). Of the 73 studies, 32 (44%) investigated the quality of the work produced by paired students, and employed various quality metrics. As can be seen in Table 2.16, quality metrics were arranged into four different categories: *Internal code quality*, *External code quality*, *Standard Quality Model*, and *General category* (McConnel, 1993). In addition, the studies were ranked (in ascending order) based on the quality metrics used the most. The table also listed the studies reported as having a significant positive effect (i.e. supporting PP), negative effect, no significant effect or providing mixed results for each quality metric.

Table 2.16 Summary of quality metrics used

Metrics' Category:	Quality Metrics(s)		Ranking (*)	Sig. +ve	Sig. –ve	No effect	Mixed
Internal Code Quality	**Program size**	1) Lines of code (LOC).	2	S53	-	S47	S21
		2) Non comment lines of code (NCLOC).	2	-	-	S25, S57	S21
		3) Comment Ratio (CR).	4	-	-	S25	
		4) Number of methods.	4	-	-		S21
	Object Oriented Design Quality	1) Method Level (McCabe's cyclomatic complexity, and number of parameters passed to the method).	3	-	-	S57	S21
		2) Class Level (Chidamber & Kemerer's metric suit).	4	S5	-	-	-
		3) Package level (Martin's package level dependency metrics).	4	-	-	S35	-
		4) Coupling Factor (CF).	4	-	-	S25	-
External Code Quality	1) Number of test cases passed/failed.		1	S33, S53, S60, S17	-	S44, S49	-
	2) Number of features correctly implemented.		4	S21	-		-
	3) Number of incomplete requirements.		4	-	-	S49	-
	4) Number of defects/errors.		2	S47, S55	-		S57, S64
	5) Completion of change request.		4	S52	-		-
	6) Thoroughness and fault finding effectiveness.		4	-	-	S36	-
	7) Number of acceptance test passed (NATP).		2	S68	-	S34, S46	-
Standard Quality Model	1) ISO/IEC 9126 Quality Factors.		4	-	-	-	S1
General	1) Programming score, project score/grade.		1	S2, S9, S30, S39, S59	-	S20,	S62
	2) Expert opinions.		1	S5, S16, S38, S73	-	S13	S21, S45

(*) The ranking showed the quality metrics used the most in PP studies (in ascending order)
(**) Some of the studies used more than one metric to measure the quality of a program/design.

Internal code quality can be divided into two sub-groups: *program size* and *Object Oriented* (OO) *design quality*. Three PP studies applied program size (e.g. LOC) as a quality metric and found that shorter programs led to higher quality and more maintainable software [S25], [S47], [S21]. However, Vanhanen and Lassenius (2005) [S57] argue that LOC is not a reliable metric because fewer lines of code does not guarantee better quality. Thus rather than using LOC as an indicator they analyzed design quality based on a method's size and

complexity metrics. There was no significant difference in performance between pair and solo students when effectiveness was measured using program size. In terms of OO design quality, program quality was rated higher for pair programmers when design quality was measured at the class level (i.e. depth of inheritance, coupling and cohesion level) [S5]. However, there was no significant difference between paired and solo students in OO design quality at the method and package levels [S21], [S35], [S57]. Hanks et al. (2004) [S21] mention that the mixture in the studies' findings was due to the various levels of task complexities assigned to the subjects. Vanhanen and Lassenius (2005) [S57] report that differences in design quality between pair and solo groups depends on the metric used and that the measures may have been affected by the size of the system analyzed.

External code quality (ECQ) was investigated by 16 (22%) of the 73 primary studies. Of these, nine (56%) report that the ECQ produced by paired students is significantly better compared with the soloist. In terms of standard quality models, only one study [S1] measured the quality of design diagrams (e.g. Data Flow Diagrams, Relational Databases, and Functional Interface Diagram), using the ISO IEC 9126 quality model. It presented mixed findings about the impact of pair work on the quality of design products [S1]. No study measured the quality of design artifacts using the Unified Modeling Language (UML) diagrams.

The general category, which comprised of expert opinion and academic performance measures such as programming score or project grade, was applied in 14 of the 73 studies (19%). Studies that relied upon expert opinion measured quality using criteria such as the significance of identifiers, how well-organized methods were, the use of appropriate indentation and whitespace [S21], functionality and style [S38], output correctness, required documentation, correct use of objects and interface design [S13], and number of defects in specification, expression, and algorithms [S45]. In 5 out of 7 studies, a program's quality produced by pairs was superior to the program's quality produced by solo students when program quality was measured using course assignment's score or project's grade [S2], [S9], [S30], [S39], [S59]. Four out of 7 studies that employed professional judges (expert opinion) to evaluate the quality of work produced by pair and solo programmers' reported that PP had a positive effect on the quality of work [S5], [S16], [S38], [S73].

In terms of the quality assessment of the studies used to answer this sub-question, the average quality score was 5.8 with the highest quality score being 7.0. Seventeen studies (55%) demonstrated good experimental quality; ten studies (32%) very good quality, and only four studies (13%) were assessed as being of fair quality.

2.3 Discussion of the SLR's Findings

In this section, a discussion of the SLR's findings is presented according to the three major areas of focus of the SLR: pair compatibility, evidence on PP's effectiveness, and quality measurement issue. We also compared the evidence with findings reported in the literature from the other research domains such as small-group research, information systems,

psychology, and educational research. Finally, the implications for research and CS/SE educators, and the threats to the validity of the SLR are also discussed.

2.3.1 Pair Compatibility Issues

The variation in studies' findings may be attributed to several factors. Some studies (e.g. [S28], [S29], [S63]) used a mixture of subjects (undergraduate and graduate students) as a representative of their population. Therefore experience and academic background may have varied widely based on students' degree level. The nature of courses, instructors, and instruments used may have also affected the studies' outcome. For instance, the instruments used in two studies that measured confidence level were different and this may have contributed to the contradictory findings [S22], [S54].

The two compatibility factors investigated most in PP studies were skill level and personality type. The synthesis of evidence suggests that pairing works effectively when students are paired according to their skill level. This evidence supports the previous work by Comrey and Staats (1955) whom report that group productivity is highly correlated with the ability or competency level of the group members.

Even though researchers speculate that personality type might have an influence on pair compatibility [S28], the studies' findings are mixed. We believe that these mixed findings could have been caused by the diversity of types of instruments used, the duration of a study, and the nature of the tasks carried out. Bowers, Pharmer, and Salas (2000), and Mohammed and Angell (2003) showed that the relationship between personality composition and team performance are highly dependent on the type of task, which supports our view.

There was very little evidence showing gender and ethnicity have an impact in relation to improving the effectiveness of PP's practice as a pedagogical tool. The issue of whether gender affected PP's effectiveness produced contradictory findings in two studies [S29], [S73]. Perhaps one of the reasons that could explain such contradictory findings was the duration of each study. The experiments carried out by Choi (2004) [S73] were conducted in a shorter duration (90 minutes) as compared with the three experiments conducted by Katira et al. (2005) [S29], each lasting for a full semester. Therefore, the effects might have been different.

We found a lack of studies investigating pair compatibility or pair effectiveness that focused on software modeling or methodologies. As our search terms included "programming" this was perhaps not an unexpected result. However, some studies applied PP to a software's design phase to investigate whether pair designing was effective at enforcing or diffusing the designs' knowledge among the project team's members [S3], [S6], [S8], [S9]. The findings indicated that pairing was beneficial in terms of knowledge transfer among pair designers. Therefore PP should not be restricted only to coding-related tasks as it may also be applicable to other software development phases.

2.3.2 Evidence on PP's Effectiveness

In total 70 studies measured PP's effectiveness using technical productivity, program/design quality, academic performance, and satisfaction criteria. Of these, 19 studies (27%) measured the pair productivity using the time spent in completing the tasks, where most findings (11 studies) indicated that pair programmers effectively completed the assigned tasks in a shorter time. One of the more significant findings to emerge from this review was that students perceived greater satisfaction and enjoyment from using PP.

From the two meta-analyses that compared PP's effectiveness against solo programming, the relatively small overall effect on final exam scores indicated that PP did not directly improve students' course grades, but the medium effect size on students' assignments scores suggests that PP was rather useful helping students in their assignments. Thus, evidence suggests that PP is an effective pedagogical tool that not only benefits students in terms of learning, but also increases their satisfaction and enjoyment. These findings corroborated the results of a meta-analysis of small group and individual learning with technology by Lou, Abrami, and d'Apollonia (2001). In particular, their synthesis of evidence suggests that students learning in pairs resulted in better cognitive and affective outcomes than those learning individually.

The cognitive theories of cooperative learning research emphasize two major benefits of students working together. First, the interaction that occurs while working together helps students increase their "mastery of critical concepts" (Slavin, 1990). When peers engage in a discussion, cognitive conflicts and reasoning are more likely to happen and this type of interaction helps improve students' achievement. Second, the ability to elaborate or explain will consequently help students in retaining the knowledge in memory. PP as one kind of cooperative learning activity clearly possesses the elements of interaction and elaboration that will help students enhance their academic achievement.

A review of research in education shows that cooperative learning can be beneficial in accelerating student' achievement when the emphasis is placed upon the *group's goals* and *individual accountability* factors (Slavin, 1980). By default, PP incorporated those factors because each pair of students are responsible for solving programming problems together whereby each student in a pair is experiencing both unique roles (i.e. as the *driver* or *navigator*). Besides, they are also accountable for their own individual achievement in exams. These are some of the factors that may explain why PP is beneficial in improving students learning.

2.3.3 Measuring Quality

The evidence gathered in the SLR supported the view that the work produced by paired students was of a high quality when measured using expert opinion and academic performance. Thirty-two studies investigated quality aspects of PP, and results in general report significant findings showing that the quality of the design/code developed by paired students was considerably superior to the quality of work of soloists. Thus, the SLR's findings

corroborated the results of the meta-analysis reported by Dyba et al. (2007).

Although results were in general supportive of PP, the effects of PP towards internal code quality seem to be unclear/contradictory. This is because most of the studies either provide a mixture of findings or report that PP had no impact on the internal code quality (see Table 2.16). For example, Madeyski (2006a) [S35] report that package dependencies in an OO design are not significantly affected by the development approach (paired or solo). Since no other evidence was, as far as we were concerned, available regarding this issue, a replication study needs to be carried out to support or refute this evidence. We believe that the unclear evidence as to whether PP improves internal code quality can be attributed to several reasons such as the type of tasks, level of task complexity, size of the analyzed system, and the studies' context.

Steiner's theories emphasized that the potential performance of a group is very much dependent on the type of task at hand, and whether the group members have adequate resources (i.e. skills, tools, and effort) in order to carry out the task (Steiner, 1972). In our SLR, the tasks given to paired students varied from simple in class programming assignments to complex J2EE distributed applications. Thus, given this range in task complexity, the internal code quality is more likely to be affected by the application size and the choice of the metrics used to measure the quality of code design. Vanhanen and Lassenius (2005) mention that measuring code design quality can be unreliable due to the varying amount of functionality offered by the application. Our review supports their findings, and we suggest that measuring quality based on external metrics (i.e. test cases passed, number of defects, scores etc.) would be a better mechanism to evaluate product/code quality. Finally, while the majority of studies investigated code quality, only three studies looked at the quality of design documents using a ISO model [S1] and/or design scores [S9], [S30].

2.3.4 Implications for Research

Based on the SLR's findings, personality was one of the most commonly investigated factors in PP studies. However, the results from existing studies are inconsistent in terms of the effects of personality on PP's effectiveness. Existing literature in psychology shows that the personality traits of students play an important role in predicting their academic success and is considered as one of the critical success factors in determining teamwork success among students (Busato, Prins, Elshout, & Hamaker, 2000; Farsides & Woodfield, 2003).

In one of the meta-analytic studies, Bowers et al. (2000) investigated whether the teams consisting of homogeneous personalities outperformed the teams consisting of heterogeneous personalities and the findings show a partial support for heterogeneous teams. While these studies were conducted mostly in the psychology domain, further research should be done in other fields too (e.g. CS/SE) in order to investigate whether personality composition can affect PP's effectiveness as a pedagogical tool. In addition, the issue of whether homogeneity or heterogeneity of personality is good for PP has not been clearly understood. We also identified that most PP studies investigated personality type

using the Myers-Briggs Type Indicator (MBTI). MBTI is one of the most popular instruments used to measure personality based on an individual preference on personality types (Murray, 1990). While MBTI is commonly used in the PP literature, we found that the Five-Factor Model (FFM) is currently considered the predominant taxonomy of personality by personality psychologists (Burch & Anderson, 2008). Therefore further research should be undertaken using other credible personality measurement frameworks such as the FFM (McCrae & John, 1992).

We also observed that in many of the PP experiments confounding errors were not controlled, leading to results that could very likely be biased (Kitchenham et al., 2002). For example, the validity of some of the results might have been confounded by the method of pair formation. For instance, instead of randomly assigning students to treatment and control, some studies let the students decide whether to pair or not (e.g. S20, S28, S63). Such optional pair arrangement might have resulted in only interested students or those who were enthusiastic about using PP to get involved in the study, thus biasing the results. In order to improve the quality of empirical research, researchers can refer to available guidelines for conducting empirical research in SE such as the one reported by Kitchenham et al. (2002). In terms of reporting controlled experiments in SE, researchers can refer to guidelines reported by Jedlitschka and Pfahl (2005).

The SLR found only 17 studies (23%) investigated factors that may affect PP's effectiveness, including pair compatibility. However, there was no clear relationship determined between pair compatibility and PP's effectiveness. Some studies investigated the perceived compatibility of students towards their partners, but no evidence was available on whether pair compatibility improved PP's effectiveness. Research in psychology has investigated the effects of interpersonal compatibility on group productivity using Schutz's FIRO theory. The results suggest that the productivity of compatible groups was greater than that of incompatible groups (Liddel, & Slocum, 1976). We suggest that the association of these factors be investigated in future PP studies.

Most of the PP studies (85%) we reviewed required students to engage in tasks only related to coding or application development, thus suggesting that PP had been rarely employed in programming courses where students were also exposed to software design/modeling tasks. This clearly indicates that further research needs to be conducted to investigate whether PP can be an effective pedagogical tool to learn CS/SE in topics other than coding. There is also a need to increase the number of studies investigating factors potentially affecting PP's effectiveness in order to aggregate results.

2.3.5 Implications for CS/SE Educators

One of the key repercussions for CS/SE educators relates to how to implement PP. The findings from the SLR suggest that the most effective pairing configuration is to pair students of similar competency levels using as a basis their exam scores/GRE/GPA or programming experience. Therefore, we suggest that educators who are willing to practice PP in their lab or

classroom should pair the students according to their skill or competency level to achieve greater pair compatibility.

The SLR also suggests that students perceive higher satisfaction when working in pairs. According to the Vygotskian theory known as "*zone of proximal development*", students are capable of achieving a higher intellectual level when collaborating with other students rather than working alone (Vygotsky, 1978). Students who pair programmed were satisfied with the pairing experience mainly because PP helped them increase their knowledge and gain greater confidence, besides improving their social interaction skills. On the other hand, PP can also assist instructors and educators in reducing their own workload as they will be a smaller number of assignments or projects to be graded.

2.3.6 Threats to the Validity of the SLR's Results

There are several factors that need to be taken into account when generalizing the results of the SLR. During the process of identifying the relevant literature we only considered as primary sources articles published electronically, thus neglecting studies that might have appeared in conference proceedings or journals that were not published online. This was particularly applicable to material published before 1987. However, since the PP practice, as considered in our SLR, was proposed in 1999 (Beck, 1999), we believe that it is less likely that PP studies are not available online. Furthermore, we used an extensive list of search databases and have included in our search all the databases we were aware of where PP primary sources had been published.

Publication bias is also considered a common issue in SLRs (Kitchenham & Charters, 2007). Publication bias is defined as the tendency to publish studies with more positive results as compared with studies producing negative results. In dealing with the issue of publication bias, we used the following strategy:

i) Develop and continuously refine the SLR protocol during the search process.

ii) Include searching of grey areas of literature such as theses, dissertations, and technical reports so that the search process covers as many studies as possible.

Another threat relates to the issue of handling the review. The main researcher was responsible for developing the protocol and carrying out the major tasks involved in each of the SLR stages, which may have unwittingly had some influence on the SLR results. However, the primary supervisor provided detailed feedback during all the stages of the SLR (e.g. protocol's preparation, primary studies' selection, data extraction's quality assurance, compiling of results), which we believe should have minimized, if not removed, any possible bias in the SLR's results. In addition, we followed very closely the recommendations suggested in the SLR guidelines (Kitchenham & Charters, 2007) in order to avoid bias.

2.4 Lessons Learned from the SLR

In this section we discuss experiences in conducting the SLR, including major issues faced during the initial phase of the SLR. These include issues on searching the relevant literature

using online databases, and challenges or difficulties in selecting studies to be included in the review. Finally, a set of recommendations for online database providers and some suggestions for future undertakings of SLRs are proposed.

2.4.1 Issues on Searching Literature Using Online Databases

The literature search is one of the fundamental tasks of a SLR process that determines the inclusion or exclusion of a study in a review (Petticrew & Roberts, 2006). Electronic or online databases are the main sources of literature used in this review. We have found that one of the difficulties in searching literature using electronic databases is the heterogeneity of search features provided by the online database providers. For example, some of the databases do not allow users to specify the search criteria so as to limit the search either on abstracts or full texts (e.g. ISI CurrentContents, SpringerLink), and there was a database that does not allow a manual construct of a search string (e.g. SpringerLink). In general each of the electronic sources we used in the SLR provides different search syntaxes and form interfaces. Similar experiences were reported by Staples and Niazi (2007) when conducting their SLR of Software Process Improvement.

Some databases (e.g. EBSCOhost, ProQuest) do not allow us to type a complete search string in a text box; instead they provide specific rows to concatenate search syntax using Boolean operators. These features are useful to users who are not familiar with constructing search using Boolean operators (such as *AND*, *OR*, *NOT* etc.). Almost all electronic databases that we searched in provided both functionalities in terms of basic (quick) and advanced searches to facilitate the search. Table 2.17 listed the comparison of online databases' features included in the SLR.

Our experience in using and searching literature from Computer Science online databases (as indexed by The University of Auckland's library) supports issues highlighted by Brereton, Kitchenham, Budgen, Turner, & Khalil (2007) who mentioned that "*Current software engineering search engines are not designed to support systematic reviews. Unlike medical researchers, software engineering researchers need to perform resource-dependent searches*" (p. 578). Thus we believe that the software engineering community should establish a link (or a one-stop portal/database) that indexes studies related to the software engineering area in order to support evidence-based research in SE. It is also important to apply a common set of interfaces in online databases in order to enable a consistency of literature search.

Apart from applying online searches, we also recommend a manual search of all volumes of conference proceedings and journals relevant to CS/SE domain. This is because some of the studies presented in conferences, seminars or workshops are not fully indexed by the digital libraries. Jorgensen and Shepperd (2007) strongly recommend manual searching of studies and suggest careful selection of journals to ensure a wide coverage of relevant studies when completeness is an issue for an SLR.

Table 2.17 Comparison of online databases features

Database name	Allow search within abstract/ full text?	Allow search by article's author?	Provide Help function/ tutorial?	Provide advanced search?	Allow manual construct of search string?	Allow limit public. Year?	Support Direct Export?	Support Import Filter?	Provide citation db?	Coverage year	Frequent of update? (daily/ monthly)
ACM Digital Library	Yes	Yes	No	Yes	Yes	Yes	No	No	No	1985-present	Daily
Current Contents	Partially(*)	Yes	No(**)	Yes	Yes (must use special field tag)	Yes	Yes	Yes	Yes	1998 -	Daily
EBSCOhost	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	1975 -	N/A
IEEExplore	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	1988 -	Daily
Web of Science	Partially(*)	Yes	No (**)	Yes	Yes	Yes	Yes	Yes	Yes	1900 -	Weekly
INSPEC	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1969 -	Monthly
ProQuest	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	1971-	N/A
Sage Full Text	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No	1960	N/A
Science Direct	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	1994-	N/A
SpringerLink	No	Yes	No	No	No	No	Yes	No	No	1997 -	N/A
Scopus	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	1966-	N/A

Note:
(*) Allow search in Topic or Title
(**) Provide only example
Note that the the *CiteSeer* and *AgileAlliances* are another two online search websites we used during the SLR, but they are not indexed by The University of Auckland collection library database.

2.4.2 Clarity of Abstracts

Selection of primary studies in an SLR primarily depends on the screening of articles' titles and abstracts. During the process of identifying relevant literature, we found that 59% (226 out of 379 studies) were irrelevant to answer the SLR's questions, leaving only 153 for final selection. Of 153 studies, we found that 52 articles (34%) showed lack of clarity in their abstract in which the information given was insufficient to help us decide whether to include or exclude the study. In such cases, the article's full text is referred to, thus delaying the selection process.

In relation to this, Budgen et. al (2007) have conducted a study regarding completeness and clarity of structured abstracts. A structured abstract is one kind of abstract written in a very objective way by explicitly stating the objective, method, results and the conclusions of the research carried out. The outcomes of the study suggest that structured abstracts are more complete and of higher clarity (easier to read) than the non-structured abstracts. We support the use of structured abstracts because they help us in identifying the aim or focus of the study in a more straightforward way.

2.5 A Comparison Between the SLR and the Existing PP's Reviews/Meta-Analyses

In this section, a set of literature review of PP research is presented to distinguish the contribution of our SLR with existing reviews and meta-analyses on PP. A SLR on the effectiveness of PP was conducted by Dyba et al. (2007) aiming at understanding the general aspects related to PP's effectiveness including "duration" (time spent to produce the system), "effort" (person-hours spent), and "quality of the final product". The review involved both professional and students as subjects and included 15 studies comparing solo and pair programming (see Table 2.18).

Table 2.18 Studies included in the Dyba's et al. (2007) meta-analysis

Author	Year	# of studies
• Nosek	1998	1
• Williams, Kessler, Cunningham, & Jeffries	2000	1
• Nawrocki & Wojciechowski	2001	1
• *****Baheti , Gehringer & Stotts • *****Rostaher & Hericko	2002	2
• Heiberg, Puus, Salumaa, & Seeba	2003	1
• Canfora, Cimitlie, & Visaggio • Muller • Vanhanen & Lassenius	2005	3
• Madeyski • Muller (2006) • Phongbaipul & Boehm • Xu & Rajlich	2006	4
• *****Canfora, Cimitile, Garcia, Piattini, & Visaggio • *****Arisholm, Gallis, Dyba, & Sjoberg	2007	2

** Studies not included in our SLR because they were conducted in an industry context*

A meta-analysis by Dyba et al. (2007) showed a positive effect size for the attributes *quality* and *duration* (0.38 and 0.40, respectively), but a negative value in terms of *effort* (-0.57). These results suggest that PP is more effective than solo programming when quality and the duration to complete the tasks are of concern. However, the negative effect on effort meant that PP required more effort (i.e. more person-hours) when compared with solo programming. This is because it takes two people to solve a task. It was also reported that the expertise and task complexity might have possibly influenced the accuracy of the studies' findings. This SLR is related to the one we carried out, but is different in terms of its purpose and population. Our SLR investigated the potential of PP as a pedagogical tool, specifically focusing on existing evidence regarding PP's effectiveness in the context of higher education institutions. The detailed comparison of similarities and differences between the two SLRs is listed below:

Similarities of this study with our SLR

- Both are using the SLR method applying the same procedure by Kitchenham & Charters (2007).
- Out of 15 studies included in Dyba's SLR, 10 of them were included in our SLR. The five studies not included were shown in Table 2.18. The reason for not including these five studies was that four of the papers involved professionals as subjects (Nosek, 1998; Rostaher & Hericko, 2002; Canfora et al., 2007; and Arisholm et al., 2007) and one investigated PP in a distributed environment (Baheti et al., 2002).
- Both SLRs performed a meta-analysis of PP's effectiveness.

Differences of this study with our SLR

- Dyba's SLR included studies both involving students and professionals, thus the context of experiments involved both industry and academia. We only focused on students as subjects, and studies conducted in higher education settings. This means that Dyba's results cannot be promptly generalized to our context since the context they used differed from ours.
- Dyba's SLR included only studies that compared the effects of PP and of solo programming. Studies that compared PP to alternative approaches were excluded. In our SLR, we included all studies that investigated factors affecting PP's effectiveness within an educational context regardless of the type of comparison (ie. pair vs solo, all pairs, PP vs side-by-side programming etc.)
- Our SLR focused on every aspect of PP that might affect students' learning – thus, issues such as satisfaction, enjoyment and confidence were considered in our SLR. In Dyba's SLR, they only focused on studies that measured the effectiveness concerning three aspects: quality, duration, and effort.
- Dyba's SLR only included formal experiments. However, we included all types of empirical studies, i.e. formal experiments, case studies, qualitative, as well as quantitative studies.

- Dyba's meta-analysis answered the question of whether PP is better than solo in terms of duration, pair effort, and software quality, whereas our meta-analysis answered the question of whether PP is better than solo in terms of students' academic achievement in their programming assignments and final exams.

In 2007, Dyba and Dingsoyr (2008a) carried out a SLR on empirical studies of Agile Software Development examining the benefits, limitations, and the strength of evidence for Agile methods. Their SLR did not specifically look at a single practice or technique related to Agile methods, such as PP, refactoring, or unit testing, but rather focused on the empirical studies about Agile software process including XP, Scrum, Crystal, DSDM, FDD, and Lean. PP was not the focus of their SLR; however, it was recognized as one of the core parts of agile methods such as XP. Based on their meta-ethnographic synthesis, they found a very low strength of evidence supporting Agile techniques, which lead to difficulties in offering specific advice to software companies and practitioners. However, the review serves as a map of findings that can be used as a basis for further investigation. Their findings also suggest the importance of undertaking more empirical studies related to Agile software development methods.

Cockburn and Williams (2001) have investigated the costs and benefits of the PP practice. They concluded that, with an increase of approximately 15% in the cost of development time, PP offers significant benefits such as improving design quality (i.e. fewer defects), rapid solutions to problems, enhancing learning process, improving team communication, and increasing the enjoyment to learn (Cockburn & Williams, 2001). They suggest that PP is a promising approach to use as a pedagogical tool and helps increase learning capacity.

A recent meta-analysis of PP's effectiveness was reported by Hannay et al. (2009). It was an extension of a SLR by Dyba et al. (2007). The earlier meta-analysis (Dyba et al., 2007) included only 15 studies, whereas the recent one (Hannay et al., 2009) included 18 studies published up and until 2007. The results of the meta-analysis showed a small significant positive overall effect of PP on *quality* when compared with solo programming. Using the fixed-effects model, the effect size 0.23 was generated, but when random-effects model is used, the effect size was 0.33. The assumptions of random-effects and fixed-effects were applied in the study due to the heterogeneity or variance in the studies' population. In terms of *duration*, a medium significant positive overall effect is reported under both random and fixed-effects models (i.e. 0.40 and 0.53 respectively). Finally, a significant negative medium overall effect was reported on the attribute "*effort*", also under both random and fixed-effects models (i.e. the overall effect of -0.73 and -0.62 are generated respectively). These results are consistent with the results in the earlier meta-analysis reported by Dyba et al. (2007).

2.6 Recent Published PP Studies Not Included in our SLR

As our SLR included studies published within 2000-2007, this section listed relevant PP studies which have not been included in our SLR. The PP empirical studies conducted in

higher education institutions and published after 2007 are summarized in Table 2.19. Four studies assessing the benefits of PP report some positive impact of PP on educational outcomes (Sison, 2008; Chigona & Pollock, 2008; Braught, Eby, & Wahls, 2008; Sison, 2009).

Table 2.19 Summary of PP studies (2008-2010)

Study	Aim(s)	Method(s)	Summary of Finding(s)
Sison (2008)	To investigate the use of PP in SE course conducted in an Asian setting.	Formal experiment using 24 students enrolled in advanced SE course.	Paired students produce programs of lower defects than that of solo students. However, no significant difference on productivity (measured by LOC/hour) is observed between the groups.
Chigona and Pollock (2008)	To study the benefits of PP for Information Systems students.	Formal experiments and surveys in introductory programming course using 32 students.	PP helps students produce program of better quality and increase students' enjoyment. There is no clear impact of PP on the knowledge transfer.
Braught et al. (2008)	To assess the effects of PP on the individual programming ability.	Formal experiment and surveys (Sample consists of 137 students).	PP helps students with lower SAT scores obtained significant improvement in individual programming skill. Students who used PP also are more likely to complete the course successfully.
Mentz et al. (2008)	To study the effects of incorporating cooperative learning principles in PP.	Quantitative and qualitative method using student teachers.	Performance is better for students who practice PP associated with cooperative learning strategies compared with students whom used PP without enforcement on cooperative learning.
Brereton et al. (2009)	To study the applicability of SLR by masters' students and to collect evidence about PP's effectiveness for teaching introductory programming.	SLR method.	PP helps improve success and retention rates of undergraduate students, increase enjoyment and confidence in learning programming; but no significant impact on exam marks and assignments. SLR method was found feasible to be applied by graduate students, with few modifications.
Cicirello (2009)	To explore students' preferences who chose to pair programmed.	Exploratory study.	Students of Math major are more likely to prefer PP compared with CS/IS majors and they prefer to pair among themselves. Students of CS/IS majors prefer to pair with non-majors.
Sison (2009)	To study the effects of PP and software size on software quality.	2 formal experiments using 48 CS students.	PP teams developed higher quality software than solo teams when solving a relatively complex task.
Braught et al. (2010)	To investigate whether pairing students by ability improve students' performance.	Post-hoc study involving 259 students.	Pairing by ability improves performance of less able students on individual programming tasks as compared with random pairing.
Salleh et al. (2009)	To study the effects of Conscientiousness on paired students' academic performance.	Formal experiment using 49 undergraduate students.	Lack of evidence for distinguishing academic performance between paired students of similar and mixed Conscientiousness.
Salleh et al. (2010a)	To study the effects of Conscientiousness on paired students' academic performance.	Formal experiment using 218 undergraduate students.	Differences in *Conscientiousness* level did not significantly affect the academic performance of paired students.
Salleh et al. (2010b)	To study the effects of Neuroticism on paired students' academic performance.	Formal experiment using 118 undergraduate students.	*Neuroticism* or lack of emotional stability did not significantly affect the academic performance of paired students.

Braught, MacCormick, and Wahls (2010) investigated the effectiveness of PP when pairing students based on the students' ability (i.e. measured by overall performance in the course). Their findings indicate that pairing students by similar ability helps improve performance of lowest quartile students on individual programming tasks, when compared with random pairing.

One study conducted a SLR of PP studies to investigate evidence about PP's effectiveness for teaching introductory programming (Brereton, Turner, & Kaur, 2009). It was a 13-week project to assess whether SLR is applicable for a taught master's degree or undergraduate students. The results indicate a promising use of this technique for collecting evidence within a limited time constraint, but the process requires attention from supervisors particularly on study selection. Brereton et al. (2009) also report that the use of PP by undergraduate students did not have significant impact on exam marks, however PP helped improve students pass and retention rates, and also confidence and enjoyment in learning programming. The results corroborate the findings from our SLR (Salleh et al., 2010).

Except for our studies (Salleh et al., 2009; Salleh, Mendes et al., 2010a; Salleh, Mendes et al., 2010b), the studies included in Table 2.19 did not change the main patterns of findings from our SLR. The findings from our studies showed that there is no significant evidence to distinguish academic performance of paired students when they are paired by personality traits Conscientiousness and Neuroticism. These studies provided additional empirical evidence regarding the effects of personality on PP's effectiveness based on the FFM.

2.7 Summary

This chapter describes the SLR process we carried out and the results gathered based on the synthesis of PP evidence in higher education settings. A total of 73 primary studies were used in the SLR, from which we identified 14 factors potentially affecting pair compatibility and/or PP's effectiveness. Of these, personality type, actual and perceived skill level were the three factors investigated the most in PP studies. We found that the results were inconsistent in terms of the effects of personality towards pairing effectiveness. PP studies that investigated actual and perceived skill levels achieved a consensus suggesting that students prefer to pair with someone of similar skills to themselves.

Evidence showed that various metrics were employed to measure PP's effectiveness, classified into *technical productivity*, *program/design quality*, *academic performance* and *satisfaction*. We found that the metric used most often to measure pair productivity was the time spent in completing the tasks. Paired students usually completed the assigned tasks in shorter duration than solo students. Almost all studies' findings reported that students' satisfaction was higher when using PP compared with working individually. In terms of academic performance, the results of meta-analysis of PP's effectiveness indicated that PP had no significant advantage in improving students' performance in final exams over solo programming (small effect size = 0.16). However, the second meta-analysis suggested that PP was effective in helping students obtaining better scores in their assignments (moderate effect size = 0.67).

There were numerous methods employed when considering quality as a measure of PP's effectiveness. Based on our review, research on quality aspects was classified into *internal and external code quality*, *standard quality model*, and *general* categories. Of all categories, external code quality and general category were the two researched the most. Findings indicated that when quality was measured according to academic performance and expert opinion, students who pair programmed produced a better quality program compared to students who programmed alone. However, when the quality of the work produced by pair and solo students was measured using metrics at the internal code level, results were contradictory.

The results of the SLR indicated that implementing PP in the classroom or lab does not lead to detrimental effects on students' academic performance. The fact that only 23% of the studies included in the review have empirically investigated factors that may affect PP's success including pair compatibility motivates further research. The remaining chapters in this thesis describe a series of formal experiments regarding the effects of personality on the effectiveness of PP in higher education context. Since existing PP studies heavily relied upon MBTI to measure personality, and MBTI has been widely criticized as a good personality test to be employed (Pittenger, 1993; Zemke, 1992), we used a personality framework based on the Five-Factor Model (FFM). The FFM, which consists of five broad personality traits (detailed in Chapter 3) is considered a robust taxonomy of personality and reported to receive the most support by personality traits researchers and psychologists (Barrick, Stewart, Neubert, & Mount, 1998; Burch & Anderson, 2008). Such a growing acceptance of FFM has motivated us to employ this framework throughout the formal experiments described in the following chapters.

Chapter 3

A REVIEW OF PERSONALITY RESEARCH AND FRAMEWORKS

This chapter introduces the Five-Factor Model (FFM) which we used as a framework for measuring personality characteristics in our research. Other major personality frameworks and models from the psychology domain are also described and relevant literature that relates personality to academic as well as team performance is summarized. Knowledge of these areas is important prior to understanding the effects of personality towards effectiveness of pair programming (PP) in a higher education context. One of the main reasons for this review is to identify the personality framework best suited to be applied in our research and to justify the rationale and motivation for selecting the framework.

3.1 Role of Personality in CS/SE Research

Research in software engineering (SE) typically involves a human element as one of its important aspects of investigation (Feldt, Angelis & Samuelsson, 2008). It has been reported, however, that there has been too much focus on the techniques, processes, and methods involved in developing software, neglecting the human issues (Feldt et al., 2008). Feldt et al. (2008) suggest that SE empirical studies should embark on gathering psychometric data on the people involved in software development. In particular, their study focussed on understanding the role of personality towards the attitude of developers to SE tools and processes. Their initial findings showed that personality traits of an individual correlate with the attitudes towards work style and adaptability to changes. Since PP is a practice that involves people working together to achieve a common set of goals, the success of the practice is largely determined by how effective they work as a team, despite their respective skills or abilities.

In this regard, numerous studies have been conducted regarding students' team performance and effective team composition based on personality traits. One major concern about team formation is to discover whether a team consisting of heterogeneous or homogeneous personalities is effective for the team's performance (Pieterse & Kourie, 2006). Rutherfoord (2001) has conducted a study using personality inventories in forming SE class projects' teams consisting of graduate students. The study's findings indicate that teams of heterogeneous personality groups outperformed those of homogeneous personality groups. It was reported that groups comprising heterogeneous personality are more open and more innovative to problem solving (Rutherfoord, 2001).

Pieterse and Kourie (2006) have investigated the role of personality within teams of tertiary students. They found that the diversity of personalities in a team had significant

positive impact on the team's success. In this study, the team's success was measured based on the team's performance (i.e. scores) on a series of project deliverables (Pieterse & Kourie, 2006). In another study, using 18 teams of students, Peslak (2006) report that the personality of team members had significant impact on project success, and diversity in team personalities did not relate to project success, thus refuting the findings reported by Pieterse and Kourie (2006). In an investigation on predictors of object oriented programming performance, Cegielski and Hall (2006) found that personality is the strongest predictor compared with cognitive ability. When it comes to performing code-review tasks, Cunha and Greathead (2007) report that people who are more intuitive perform better than those who are less intuitive. They also suggest that companies should capitalize on the strengths of workers based on their personality in order to improve productivity.

The strategies for effective software project team formation or composition based on personality have been investigated in several research projects involving professionals (e.g. Gorla & Lam, 2004; Bradley & Hebert, 1997). Bradley and Hebert (1997) suggest that a team composed of heterogeneous or diverse personalities is more capable of performing better, thus increasing team productivity. However, Gorla and Lam (2004) argue that there is no significant effect of member heterogeneity for a small team size due to team members involvement in multiple stages of a software development process.

In the PP literature, evidence from our Systematic Literature Review (SLR) showed that personality is one of the factors most often investigated in PP studies using the Myers-Briggs Type Indicator (MBTI) as personality measurement (Salleh et al., 2010). Table 3.1 summarizes the existing PP studies conducted in academic and industrial settings that investigated the impact of personality on PP. In general, findings from these studies were quite diverse, and thus inconclusive, on whether personality could significantly affect the outcome or productivity of pair programmers. Only two studies (Sfetsos et al., 2009; Choi et al., 2008) presented positive findings, reporting that paired students of different personality types performed better when compared with paired students of similar personality types. Most studies reported that personality had no significant influence in determining PP's effectiveness (Chao & Atli, 2006; Heiberg et al., 2003; Katira et al., 2005; Gevaert, 2007).

To date, empirical findings using the FFM report low support for the effects of personality in PP. For instance, Hannay et al. (2010) report personality as only a moderate predictor for pair performance. They suggest that the performance of pair programmers may also be affected by other factors such as expertise, and task complexity. Personality traits Conscientiousness and Neuroticism were reported not to significantly affect paired students' academic performance (Salleh, Mendes et al., 2010a; Salleh, Mendes et al., 2010b). Other empirical study reported by Acuna et al. (2009) investigated the relationship between personality, team processes, task characteristics, software quality and team's satisfaction in students' team practicing Agile XP methodology. Their findings indicate that personality factor Extraversion is positively correlated with software quality, and teams with higher aggregate on Agreeableness and Conscientiousness achieved the highest job satisfaction.

Table 3.1 List of PP studies investigating personality factor

Author(s)	Type of study	Sub.	Size	IV	DV	Outcomes	Personality measurement
Chao & Atli (2006)	Survey & Exp.	Stud.	58	Personality traits	PP success (code quality and pair compatibility)	PP success is not influenced by differences in personality traits.	Personality characteristics (Univ. of Denver Career Centre)
Heiberg et al. (2003)	Formal Exp.	Stud.	110	PP vs. Non-PP	PP productivity	The individual personality traits do not have significant consequences to PP performance.	NEO PI
Katira et al. (2004)	Formal Exp.	Stud.	564	Personality, skill level, technical competence, and self-esteem	Pair compatibility	Results were mixed. Personality differences affect compatibility of freshmen but not for advanced undergraduate students.	MBTI
Katira et al. (2005)	Formal Exp.	Stud.	361	Personality, skill level, self esteem, gender and ethnicity	Pair compatibility	Pair compatibility was not affected by personality of the paired students.	MBTI
Layman (2006)	Survey	Stud.	119	All paired	Perception towards collaboration	Personality had no significant effect towards perception to collaborate.	MBTI
Sfetsos et al. (2009)	Formal Exp.	Stud.	70	Personality	PP's effectiveness	Paired of mixed personalities performed better than paired of same personality.	KTS
Williams et al. (2006)	Formal Exp.	Stud.	1350	Personality, learning style, skills, self esteem, work ethic	Pair compatibility	Results were mixed. Partial supports of personality in predicting compatibility.	MBTI
Choi et al. (2008)	Formal Exp.	Stud.	128	Personality	PP outcome (code productivity)	Personality differences have significant impact on PP outcomes.	MBTI
Gevaert (2007)	Formal Exp.	Stud.	28	PP Vs Solo	Time spent	Personality does not significantly affect the efficiency of students who paired	Eysenck Personality Questionnaire
Dick & Zarnett (2002)	Case studies	Prof.	8	N/A	N/A	Personality traits critical for PP success were communication, comfortableness working in a team, confidence and ability to compromise.	N/A
Hannay et al. (2010)	Regres sion	Prof.	196	Personality	Pair performance	The effects of personality were not consistent and suggest that personality as only a moderate predictor for pair performance.	Big Five
Salleh et al. (2009)	Formal Exp.	Stud.	54	Personality trait Conscientious ness	PP's effectiveness	Heterogeneity of personality trait Conscientiousness had no significant effect on paired students' performance.	Five-Factor Model
Salleh et al. (2010a)	Formal Exp.	Stud.	218	Conscientious-ness	PP's effectiveness	Differences in Conscientiousness level did not significantly affect paired students' academic performance	Five-Factor Model
Salleh et al. (2010b)	Formal Exp.	Stud.	118	Neuroticism	PP's effectiveness	Paired students' performance was not significantly affected by the different levels of Neuroticism.	Five-Factor Model

Exp – Experiment Sub. – Subject Stud. – Student Unk. – Unknown N/A – Not available Prof. – Professional
IV – Independent Variable DV – Dependent Variable

3.2 Major Personality Theories

In studying personality, six major types of personality theories have been developed over the past 100 years (Burger, 1993): a) Psychoanalytic, b) Trait, c) Biological, d) Humanistic, e) Behavioral/social learning, and f) Cognitive. The *psychoanalytic* theory was proposed based on behavioral observations made by Sigmund Freud in the late 1800s. This theory emphasizes the concept of the conscious-unconscious mind in understanding human personality. Other theorists associated with these concepts were Carl Jung who later invented the theory of psychological types, and Alfred Adler who consider individual personality as a hereditary behavior (Kasschau, 1985).

The *trait* theory, which was initially influenced by the work of Gordon Allport, Raymond Cattell, and Henry Murray (Burger, 1993), is recognized as the most widely accepted approach in describing and predicting behavior (Burch & Anderson, 2008). Factor analytic studies of human characteristics or traits produced a set of personality dimensions known as the "Big Five" or Five-Factor (Digman, 1990). Each personality trait is associated to human's behavior, represented by human responses to a specific situation (Digman, 1990). For example, the Agreeableness trait represents a personality dimension that involves the more humane aspects of humanity; the characteristics such as altruism, nurturance, compliance, and tender-mindedness describe one end of the dimension, whereas hostility, self-centeredness, suspicious, and distrust describe the other end (Driskell et al., 2006; Digman, 1990).

The *biological* approach considers genetic influence and physiological processes in describing personality; whereas the *humanistic* approach identifies differences in behavior based on personal responsibility and individual perception towards the environment (Burger, 1993). According to the *behavioral/social learning* approach, personality is composed of individual learning experiences, gathered through observations of other people's behaviors (Burger, 1993). Finally, the *cognitive theory* emphasizes that cognitive structures or the way people process information explains individual differences in personality.

The following subsections briefly discuss the five major personality frameworks most often used in the research in the computing and personality psychology domains (Feldt et al., 2008; Digman, 1990; Peslak, 2006; Hannay et al., 2010) : The Five-Factor Model, Myers-Briggs Type Indicator, Keirsey Temperament Sorter, Cattel's 16 Personality Factor, and Eysenck Personality.

3.2.1 The Five-Factor Model (FFM)

The Five-Factor Model (FFM), also known as the "Big Five" (see Figure 3.1), is a taxonomy of personality comprising of five broad personality traits - Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism – and provides a structure that categorizes dimensions of differences in human personality (McCrae & John, 1992). This model was derived using factor analytic research based on *trait* theory. Factor analytic

research refers to multiple studies that analyse the comprehensive set of natural-language terms used to describe an individual's personality, where replication of the studies had identified the five clusters of traits (John & Srivastava, 1999). As mentioned by John and Srivastava (1999), "*the Big Five structure does not imply that personality differences can be reduced to only five traits. Rather, these five dimensions represent one's personality at the broadest level of abstraction, and each dimension summarizes a large number of distinct, more specific personality characteristics*" (p. 105). Each of the five traits is discussed below:

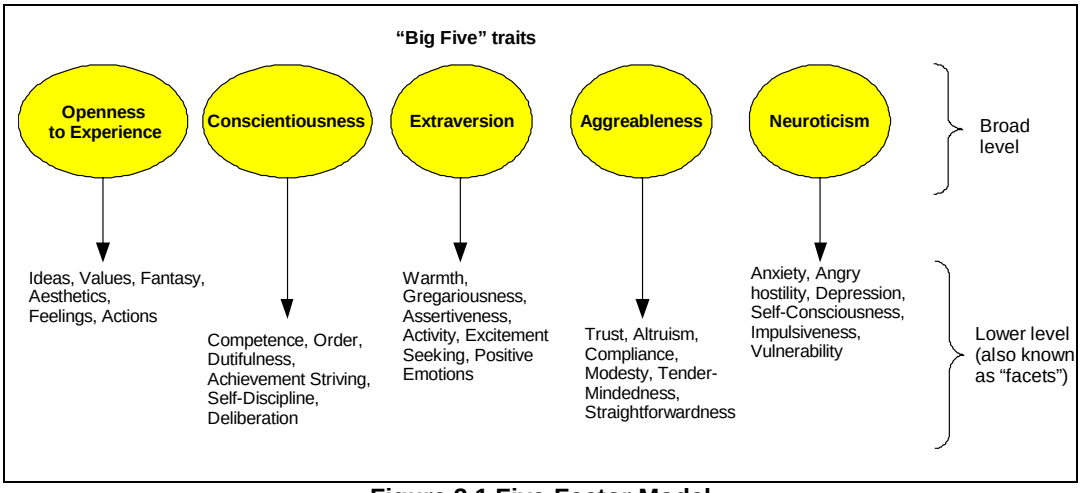


Figure 3.1 Five-Factor Model

a) Openness to Experience

Openness to experience describes intellectual, cultural, or creative interest (Driskell et al., 2006). Someone who is high on Openness to experience tends to appear as imaginative, broad-mindeded, and curious, whereas those at the opposite end of this spectrum usually show a lack of aesthetic sensibilities, preference for routine, and favouring conservative values (Barrick & Mount, 1991).

b) Conscientiousness

Conscientiousness is concerned with one's achievement orientation. This trait has been consistently reported to be related to work performance (e.g. Barrick et al., 1998; Witt et al., 2002; English et al., 2004). People who are high in Conscientiousness tend to be hardworking, organized, able to complete tasks thoroughly and on-time, and also reliable. On the other hand, low Conscientiousness relates to negative traits such as being irresponsible, impulsive, and disordered (Driskell et al., 2006).

c) Extraversion

Extraversion relates to the degree of sociability, gregariousness, assertiveness, talkativeness, and activeness (Barrick & Mount, 1991). A person is considered an extravert if he/she feels

comfortable in a social relationship, friendly, assertive, active, and outgoing. It was reported that extravert members are expected to stimulate group discussion, but the inclusion of too many extraverts could be destructive to the team (Mohammed & Angell, 2003; Neuman et al., 1999). Thus an intermediate level of Extraversion in a team leads to better performance (Barry & Stewart, 1997).

d) Agreeableness

Agreeableness refers to positive traits such as cooperativeness; kindness, trust and warmth, and persons low on Agreeableness tend to be sceptical, selfish, and hostile. A team that requires a high level of collaboration or cooperation can benefit from agreeable team members. Witt et al. (2002) report that there is a positive link between both Agreeableness and Conscientiousness traits. Their findings suggest that conscientious workers with high level of Agreeableness receive higher job performance ratings.

e) Neuroticism

Neuroticism refers to the state of emotional stability. Someone low in Neuroticism tends to appear calm, confident, and secure, whereas a high Neuroticism individual tends to be moody, anxious, nervous, and insecure (Driskell et al., 2006). In a review of personality in education, De Raad & Schouwenberg (1996) state "*particularly at the University level, highly neurotic students are probably handicapped as compared to low neurotics*" (p. 326). Thus this factor is believed to affect academic performance. Neuroticism is also reported to be consistently related to self-efficacy and the relationship between Neuroticism and self-efficacy is moderated by gender (Schmitt, 2008).

According to John and Srivastava (1999), the Big Five dimensions represent human personality at a broad level and were derived based on the hierarchy of personality descriptors. At the lower level of the hierarchy, these factors can be narrowed down into what is known as "facets" (Costa & McCrae, 1995). Figure 3.1 shows the 30 facet scales of NEO-PI-R's inventory, which have been identified and empirically validated by Costa & McCrae (John & Srivastava, 1999; Costa & McCrae, 1995).

In terms of operationalizing the Five-Factor Model, there are various instruments developed to measure personality using the big-five traits. *NEO Five-Factor Inventory* (NEO-FFI) is one of the instruments that is well accepted, widely assessed, and extensively used to measure the Big Five personality dimensions (Matzler et al., 2008; Chamorro-Premuzic & Furnham, 2003a; Farsides & Woodfield, 2003; Conard, 2006). The *Revised NEO Personality Inventory* (NEO-PI-R) is also another well-established instrument developed by Costa & McCrae (1992a) to measure 30 personality facets. In recent years, a Web-based instrument known as *International Personality Item Pool* (IPIP) was developed by Goldberg and Johnson (Goldberg, 1999; Goldberg et al., 2006). While the NEO-FFI and the NEO-PI-R are proprietary instruments, the IPIP is freely accessible in the public domain website (Goldberg et al., 2006). IPIP was developed by personality psychologists via an international collaboration of research for the purpose of providing an inventory that is available for

comparative validation (Goldberg, 1999). Such validation helps improve the reliability of the inventory as a personality measurement scale (Gow et al., 2005). It has also been reported that such an automated instrument (i.e. computer-based) is much more efficient compared with any paper-based personality instruments (Goldberg et al., 2006).

3.2.2 Myers-Briggs Type Indicator (MBTI)

In contrast to the FFM, the Myers-Briggs Type Indicator (MBTI) is designed to measure a preference of personality types (Furnham, 1996). The MBTI instrument, which was developed by Katharine Cook Briggs and Isabel Briggs Myers during World War II, was based on the personality type theory of Carl Jung (Myers & Myers, 1995). In defining preferences, Myers and Myers (1995) categorize individual behavior into different cognitive functions represented by the following four dichotomies of preferences:

- Extraversion (E) versus Introversion (I).
- Sensing (S) versus Intuition (N).
- Thinking (T) versus Feeling (F).
- Judging (J) versus Perceiving (P).

The *extraversion and introversion* preferences describe how people respond to the outer world (i.e. outward/inward direction toward people and objects). The extravert is usually active and energetic when dealing with the outside world, as opposed to the introvert who prefers to indulge in their own thoughts and ideas. The *sensing and intuition* dimension refers to the way people perceive or gather information, whereas the *thinking and feeling* dimension reflects how people make a decision. The sensor (S) individuals perceive information based on facts, or precise data; intuitive (I) individuals tend to make associations among data and discover possible alternative patterns of information. The thinker (T) individuals tend to be more logical and analytical when making decisions, whereas the feelers (F) consider personal values and empathize with the situations when rationalizing evidence. The Judging (J) type prefers a careful and organized plan in dealing with the outside world, but the perceiving (P) type is rather spontaneous, flexible, and open (Myers et al., 1998). The combination of personality types results in 16 sets of personality preferences, as can be seen in Table 3.2 (Myers & Myers, 1995).

Table 3.2 The 16 MBTI types

ISTJ	ISFJ	INFJ	INTJ
ISTP	ISFP	INFP	INTP
ESTP	ESFP	ENFP	ENTP
ESTJ	ESFJ	ENFJ	ENTJ

3.2.3 Keirsey Temperament Sorter (KTS)

The Keirsey Temperament Sorter (KTS) instrument was proposed by Keirsey and Bates (1984) and later revised by Keirsey in 1998 (Keirsey, 1998). It was based on the theory of

temperament types and also a simplified version of the MBTI test (Keirsey & Bates, 1984). Instead of classifying personality types into sixteen slots, KTS combined the MBTI's sensing and perceiving functions, and the intuitions with the judging functions, thereby generating four temperament types (Keirsey, 1998):

- *Artisan* (seeking stimulation/inspiration, and virtuosity, concerned with *making an impact*)
- *Guardian* (seeking security and belonging, emphasize on responsibility and duty)
- *Idealist* (seeking meaning and significance, are intuitive and cooperative)
- *Rational* (seeking mastery and self control, typically intuitive, practical and realistic)

According to Keirsey (1998), individual temperaments can be described as a form of rings consisting of the inner ring and the outer rings. The inner ring represents the perceiving functions of a human being, or the way people receive and process information. Keirsey categorizes the perceiving function as either an *abstract (intuitive)* or *concrete (sensing)* type. The second ring represents the judging functions, concerned with how people make decisions and rationalize their judgments (Keirsey, 1998). The two types of judging functions are *cooperative* and *pragmatic*. The combination of the inner ring and the second ring determines an individual's temperament type as shown above. The third ring represents human attitude towards the outer world or the way people communicate. The two types of attitude are *directive* and *informative*. These attitudes distinguish the type of role played by each of the temperaments. Finally, the fourth ring represents the way people get their energy, either from the outside world or from the inner world. The two types of orientation are known as *expressive* (extravert) or *attentive* (introvert). Table 3.3 shows the relation of rings to one another and to different types of temperament that describe human personality.

Table 3.3 The Keirsey Temperament (Keirsey, 1998)

	Temperament	**Role**	**Role Variant**
Inner Ring	*2nd Ring*	*3rd Ring*	*4th Ring*
Abstract Vs Concrete	**Cooperative Vs Pragmatic**	**Directive Vs Informative**	**Extravert Vs Introvert**
Abstract (N)	**Idealist** (NF)	Mentor (NFJ)	Teacher (ENFJ) Counselor (INFJ)
		Advocate (NFP)	Champion (ENFP) Healer (INFP)
	Rational (NT)	Coordinator (NTJ)	Fieldmarshal (ENTJ) Mastermind (INTJ)
		Engineer (NTP)	Inventor (ENTP) Architect (INTP)
Concrete (S)	**Guardian** (SJ)	Administrator (STJ)	Supervisor (ESTJ) Inspector (ISTJ)
		Conservator (SFJ)	Provider (ESFJ) Protector (ISFJ)
	Artisan (SP)	Operator (STP)	Promoter (ESTP) Crafter (ISTP)
		Entertainer (SFP)	Performer (ESFP) Composer (ISFP)

Note: N – Intuitive; S – Sensing; F – Feelers; T – Thinker
J – Judging; P – Perceiving; E – Extrovert; I – Introvert

The major difference between the Myers Briggs types and Keirsey's temperaments is that the former are concerned with how people think and feel, whereas the latter are concerned

with directly observable behaviors (Francis, Craig, & Robbins, 2008). In terms of classifying the personality type, MBTI put emphasis on the extraversion and introversion (i.e. expressive/attentive) dichotomy, whereas KTS stresses on the importance of the sensing/intuition (i.e. concrete/abstract) perspective (Francis et al., 2008).

3.2.4 The Sixteen Personality Factors (Cattell's 16PF)

The Cattell's System (1905-present) was derived based on factor analytic studies of peer ratings of college students by the psychologist Raymond Cattell (Burger, 1993). Cattell's work aimed to discover the number of basic personality traits required in order to understand the structure of human personality. His analysis resulted into the 16 personality factors (16PF) listed in Table 3.4, which can be used to describe individual personality traits (Burger, 1993). This personality model is considered complex and independent replication showed that correlations of the 16PF are reported to generate not more than seven factors (Digman, 1990). Cattell's contributions, however, were recognized by Digman (1990) as "*important and essential for the development of a quantitative approach to personality assessment*" (p. 428).

Table 3.4 Cattell's 16 Personality Factor (Burger, 1993; Conn & Rieke, 1994)

Factor	Low Score Description	High Score Description
Warmth	Reserved, detached, formal	Warm, outgoing, participating
Reasoning	Concrete-thinking, lower general mental ability	Abstract-thinking, more intelligent
Emotional Stability	Affected by feelings, emotionally less stable	Mature, emotionally stable
Dominance	Coooperative, submissive, accommodating	Dominant, assertive, competitive
Liveliness	Serious, restrained, introspective	Lively, enthusiastic, cheerful
Rule-Conscientiousness	Expedient, nonconforming, disregards rules	Rule-conscious, dutiful, conforming
Social Boldness	Shy, threat-sensitive, timid	Socially bold, venturesome, uninhibited
Sensitivity	Tough-minded, self-reliant, objective	Tender-minded, sensitive, clinging
Vigilance	Trusting, accepting, unsuspecting	Vigilant, suspicious, distrustful
Abstractness	Practical, grounded, conventional	Abstract, imaginative, absentminded
Privateness	Forthright, unpretentious, genuine, open	Private, discreet, astute
Apprehension	Self-assured, secure, complacent	Apprehensive, self doubting, worried
Openness to Change	Traditional, conservative	Experimenting, liberal
Self-Reliance	Group-dependent, a "joiner" and sound follower	Self-reliant, solitary, prefers own decision
Perfectionism	Tolerates disorder, impulsive, uncontrolled	Controlled, perfectionists, organized
Tension	Relaxed, tranquil, patient	Tense, frustrated, impatient

3.2.5 Eysenck Personality

Hans Eysenck discovered the personality system based on factor analytic research of biological traits (Burger, 1993). Eysenck initially included only two basic dimensions of personality: *Extraversion-Introversion* and *Neuroticism* (Burger, 1993). His further research identified the third dimension known as *psychoticism* (Eysenck & Eysenck, 1975). The Extraversion-Introversion and Neuroticism dimensions are similar to the FFM traits, whereas psychoticism is reported to represent a dimension consisting of a combination of the Agreeableness and Conscientiousness traits (Burger, 1993; Digman, 1990).

According to Eysenck, the level of extraversion is determined based on the variability of cerebral cortex arousal (Burger, 1993). For instance, extraverts generally possess a lower level of cortical arousal compared with the introverts. Thus, extraverts tend to seek for external stimulation such as being outgoing, socializing and getting involved in group activities. On the other hand, introverts' cortical arousal operates at an above-optimal level, thereby making them more solitude and non-stimulate to the environment. In terms of the Neuroticism dimension, high scores indicate high levels of negative effects such as depression and anxiety. Thus, people who score low on this dimension are considered emotionally stable. The psychoticism dimension comprises traits ranging from tendermindedness through toughmindedness, and including psychotic disorder characteristics (Burger, 1993). Table 3.5 lists the traits associated with the Eysenck's personality model. Eysenck Personality can be measured by using the *Eysenck Personality Questionnaire (EPQ)*, or more recently the *Revised Eysenck Personality Questionnaire (REPQ)* (Francis et al., 2008).

Table 3.5 Eysenck's personality (Wikipedia, 2010)

Psychoticism	**Extraversion**	**Neuroticism**
Aggressive	Sociable	Anxious
Assertive	Irresponsible	Depressed
Egocentric	Dominant	Guilt Feelings
Unsympathetic	Lack of reflection	Low self-esteem
Manipulative	Sensation-seeking	Tense
Achievement-oriented	Impulsive	Moody
Dogmatic	Risk-taking	Hypochondriac
Masculine	Expressive	Lack of autonomy
Tough-minded	Active	Obsessive

The following subsections detail the motivation and rationale for selection of the personality model used in this research. A review of literature from relevant domains (e.g. psychology and business) regarding the impact of personality towards teams' effectiveness and academic performance is also discussed.

3.3 Motivation/Rationale for Using the FFM

Our systematic review showed that most of the existing PP research on personality employed the MBTI as an instrument to measure personality type (Salleh et al., 2010). MBTI is reported as one of the most popular instruments to measure an individual's personality for non-psychiatric populations, and also has been used extensively in the business domain (Murray, 1990). Also, most of the research in the Information Systems and CS/SE domains typically uses MBTI for the purpose of investigating the effects of personality towards the performance of students' team projects (e.g. Peslak, 2006; Karn & Cowling, 2006; Capretz, 2002; Bradley & Hebert, 1997).

Even though MBTI is very popular and widely used, there has been some criticism about the reliability and the validity of this instrument (e.g. Feldt et al., 2008; Davito, 1985; McCrae & Costa, 1989; Schriesheim et al., 1991). The MBTI test is criticised in regard to its reliability and validity as a personality measurement test (Davito, 1985; McCrae & Costa, 1989; Schriesheim et al., 1991), in particular for not having a bimodal distribution in terms of its statistical structure. As a result, any data distortion can cause serious psychometric shortcomings (McCrae & Costa, 1989; Schriesheim et al., 1991).

In spite of MBTI's popularity, we found that the Five-Factor personality model is currently considered the predominant taxonomy of personality by personality psychologists (Burch & Anderson, 2008; Furnham, 1996; Costa & McCrae, 1995). Therefore, in our research, an instrument based on the Five-Factor Model (FFM) is employed. Our selection of FFM as a personality assessment framework was due to its comprehensive nature and its ability to capture the basic temperament and dispositional factors relevant to the educational context (De Raad & Schouwenburg, 1996). Besides, there is a growing consensus among personality trait researchers that FFM consists of a robust taxonomy of personality (Farsides & Woodfield, 2003; Neuman et al., 1999). In terms of its validity and reliability, FFM is generally accepted by personality psychologists who suggest that such a broad trait of dimensions adequately represents human personality attributes (Barrick & Mount, 1991; Barrick et al., 1998).

In comparison with MBTI, FFM was derived based on research on the classic trait theory, whereas MBTI was developed based on Jung's theory of psychological types (Furnham, 1996). In terms of the scoring method used to measure personality, MBTI classifies an individual's personality into 1 of 16 different personality types using the combination of the four dichotomous preferences. In FFM, using a five-point Likert scale, the scoring is made by summing the numerical scores of each facet's part of the factor. The scores for each factor are represented in numerical scales with zero (0) being the lowest score, and 99 the highest score (Johnson, 2008). Thus, MBTI uses a bipolar discontinuous scale, in contrast to the continuous scale used by the FFM. The quantitative nature of the FFM scale allows us to perform more powerful statistical testing (i.e. parametric tests) when compared with those of non-parametric statistical tests, which would need to be employed with other personality frameworks due to the measurement scales they employ (Feldt et al., 2008).

3.4 Review of Research on Personality and Team Composition

Team composition refers to a process of arranging a team based on its members' attributes such as personality, demographics and other individual characteristics of team members (Levine & Moreland, 1990). Evidence from existing research suggests that team composition has a significant influence on team performance and thus can be useful for organizational restructuring, selection for team-based jobs, and selection into teams (Bell, 2007). Understanding the theories proposed in other domains (e.g. psychology) on the issue of composing a successful team can be beneficial for CS/SE education, in particular to improve the pair formation approach of PP teams. This knowledge is useful for understanding ways to improve performance of CS/SE students' team and students' academic performance in general.

Neuman, Wagner and Christiansen (1999) have investigated the relationship between team effectiveness and team personality composition in two different aspects: *Team Personality Elevation* (TPE), and *Team Personality Diversity* (TPD). They also proposed two models for determining whether a heterogeneous or homogeneous team is better or preferable for improving team performance: The *supplementary* model suggests that team homogeneity is positively related to team success, whereas the *complementary* model suggests that performance is improved when personalities among team members are diverse or heterogeneous (Neuman et al., 1999).

A summary of literature from the psychology, education, and computing domains on the relationship between personality and team composition, and the influence of personality on team performance are presented in Table 3.6. In determining an effective personality composition, some studies suggest that the type of task is an important factor that can influence team performance (e.g. Mohammed & Angell, 2003; Neuman et al., 1999; Bowers et al., 2000; Peeters et al., 2006), because the tasks performed by a team strongly determine the kind of composition likely to affect team effectiveness (Mohammed & Angell, 2003). In this regard, evidence from the research suggests that the clear advantages or benefits of homogeneity or heterogeneity on any attributes or traits cannot be ascertained or concluded due to their dependency on the nature of a task (Bowers et al., 2000).

In measuring personality, there is a clear distinction between the personality models applied across different domains: the personality types of MBTI are commonly used in Information System (IS) research, but FFM is more favoured among personality-psychology researchers and clinical psychologists. Note that some literature uses the term "Big Five" but they are referring to the same five factors as in the FFM framework. Therefore, for consistency we will use the term "FFM" throughout this thesis.

Table 3.6 A summary of the literature review on the relationship between personality and team composition

Author(s) & Year	Aim	Findings	Team type/ personality model(s)	Support Heterogeneous teams?
Peslak (2006)	To investigate the role of personality in relation to Information Technology (IT) team project processes and team success.	The study does not find any relationship between personality and team processes (i.e. team roles/team building, leadership, communication). Personalities have a significant impact on overall team success (measured by project scores). Project success improved with higher levels of Extraversion, thinking, and judging. Team diversity was not found to significantly influence overall IT project success.	18 student teams of 2 to 5 members (Used MBTI).	Did not support heterogeneous teams.
Karn & Cowling (2006)	To investigate the effects of personality on the performance of SE teams.	Results suggest that a team of heterogeneous personality and ethnicity managed to work efficiently. Project team should discuss their work and lack of debate of or disruption to the debate will affect team performance.	3 student teams of 5 members (Used MBTI).	Supports for heterogeneous teams.
Pieterse & Kourie (2006)	To investigate the role of personality diversity towards SE team performance of tertiary students.	Personality diversity had positive impacts on team performance. There was strong correlation between personality diversity in team and the team success.	Student teams of 4 or 5 members (Used KTS).	Support for heterogeneous teams.
Bradley & Hebert (1997)	To investigate the impact of personality type on the productivity of IS development teams.	Personality types are an important factor that can affect team performance. Results suggest that a team consists of diverse and balance personality types are more successful.	2 teams of IS professionals (Used MBTI).	Support for heterogeneous teams.
Neuman et al. (1999)	To investigate team effectiveness based on two aspects of work-team personality composition: Team Personality Diversity (TPD) & Team Personality Elevation (TPE).	When TPD is considered, team performance can be predicted based on *Extraversion* and *Emotional Stability*. For TPE, job performance of the team can be predicted by the *Conscientiousness*, *Agreeableness* and *Openness.* Results suggest that heterogeneous teams performed better than homogeneous group; however, these depend on the type of task performed by the team.	Team of employees (Used Big Five).	Support for Heterogeneous teams.
Mohammed & Angell (2003)	To investigate the effect of personality heterogeneity on team performance.	Relationships between personality composition and team performance are highly dependent on the type of task. The study found that higher variability on Agreeableness and Neuroticism resulted in lower performance of oral task, whereas teams with higher variance on Extraversion performed better in oral tasks. Overall results suggest that personality composition influences team performance, but depends on the nature of task.	59 student teams. Used FFM (NEO-FFI).	Does not support any form.

Barrick et al. (1998)	To examine relationships among team composition (personality & ability), team process, and team outcomes.	Results supported the hypotheses that greater general mental ability contributes to team success. Teams with higher levels of Conscientiousness, Agreeableness, and Emotional Stability achieved better performance on additive tasks. Results also strongly supported that teams with higher mean levels of Extraversion and Emotional Stability were associated with team viability (capability to continue working together).	Team of employees. Used FFM (Personality Characteristics Inventory).	N/A.
Vianen & Dreu (2001)	To examine the relationships between personality in teams, task cohesion and team performance.	Conscientiousness and Agreeableness correlated positively to both task cohesion and team performance, whereas Extraversion and Emotional Stability contributed positively to social cohesion. Task characteristic is identified as a factor influencing the relationships.	Used Five-Factor Personality Inventory (FFPI).	Support for Homogeneity in personality within teams.
Bowers et al. (2000)	To perform a meta-analysis on studies that investigates the effects of homogeneity or heterogeneity of team members on team performance with respect to gender, ability, and personality.	The integration of studies showed no clear advantage of homogeneity or heterogeneity of particular attributes towards team performance. Team success was dependent upon the type of tasks. Results suggested that homogeneity of a group had very little effect on task performance, particularly on low-difficulty tasks.	N/A.	Does not support any form.
Peeters et al. (2006)	To perform a meta-analysis on the relationship between team composition and team performance focusing on the trait elevation and variability of personality traits.	In terms of trait elevation, teams with higher average level of *Agreeableness* and *Conscientiousness* achieve better performance (effects are more salient for professional teams). Different effects were observed between professional and students teams for *emotional stability* and *Openness to experience* factors.	Used FFM.	Support for homogeneous teams.
Kichuk & Wiesner (1997)	To examine the relationships between personality and performance of product design teams.	Some of the traits could be predictive of team performance. Performance increased for teams consisting of higher level of general cognitive ability, higher *Extraversion*, higher *Agreeableness* and lower *Neuroticism*. Team performance is lowered by heterogeneity of Conscientiousness among team members.	419 first year engineering students (Used NEO-FFI).	Support for homogeneity in terms of Conscientiousness.
Bell (2007)	To investigate the relationships between deep-level team composition variables and team performance using meta-analysis procedure.	The meta-analysis shows that several traits were related to team performance: medium effect size for Agreeableness and Conscientiousness; small effect size observed for emotional stability, Extraversion, and Openness to experience. The relationships were strongly moderated by the type of settings where the studies were conducted. Analysis shows that the effects were more prominent in a field setting than in the lab setting.	Used FFM.	Low support for heterogeneity (in terms of Extraversion).

3.5 Review of Research on Personality and Academic Performance

The issue of personality in predicting academic performance (AP) has been researched in various studies (Poropat, 2009; De Raad & Schouwenburg, 1996; Chamorro-Premuzic & Furnham, 2008). The relationship between personality traits (using the big-five theory) and academic performance is summarized in Table 3.7. Our review of research on the relationship between personality and academic performance (see Table 3.7) showed that various instruments have been employed in order to measure personality (e.g. 5PFT, IPIP, NEO-FI, NEO-PI-R). Of 15 studies, 14 studies (Busato et al., 2000; Duff et al., 2004; Pulford & Sohal, 2006; Komarraju et al., 2009; Chamorro-Premuzic & Furnham, 2008; Chamorro-Premuzic & Furnham, 2003a; Furnham et al., 2003; Chamorro-Premuzic & Furnham, 2003b; Dollinger & Orf, 1991; Lounsbury et al., 2003; Nguyen et al., 2005; Fruyt & Mervielde, 1996; Conard, 2006; Poropat, 2009) report that Conscientiousness is significantly positively associated with academic performance. In these studies, academic performance is measured by various indicators including exam grades, exam scores, GPA, course grades, and mid term test.

An exploration of primary traits taking into account the facets within each trait discovered that dutifulness and achievement striving are the two important facets of Conscientiousness strongly associated with academic success (Chamorro-Premuzic & Furnham, 2003b). Consistent with these findings, the results from a meta-analysis related to personality and academic performance also report the significant positive association of Conscientiousness with AP in tertiary education (Poropat, 2009). The meta-analysis also reports positive correlations for Agreeableness and Openness, and suggests that personality-based FFM can be a valid predictor of AP, similarly to intelligence tests (Poropat, 2009).

Of the five traits, Neuroticism is reported to negatively correlate with academic performance in two studies (Chamorro-Premuzic & Furnham, 2003a; Chamorro-Premuzic & Furnham, 2003b). These findings however, were contradicted by the results reported in (Komarraju et al., 2009), which show that well-performing students also experience certain degrees of Neuroticism (i.e. feeling worried and anxious). In relation to personality traits, some studies investigate other education-related factors relevant in influencing academic performance of students. These include learning styles (e.g. Busato et al., 2000; Duff et al., 2004; Chamorro-Premuzic & Furnham, 2008), motivation (e.g. Busato et al., 2000; Komarraju et al., 2009), cognitive ability (Furnham et al., 2003), and intelligence level (e.g. Chamorro-Premuzic & Furnham, 2008; Furnham et al., 2003). The findings reported by Komarraju et al. (2009) and Busato et al. (2000) highlight significant relationships between academic success, personality traits and motivation. In these studies, highly motivated students showed greater academic achievement and a positive link between academic motivation and Conscientiousness trait suggests that motivated students are also conscientious and/or organized.

Table 3.7 Summary of the literature review on the relationship between personality traits and academic performance (based on FFM)

Author(s) & Year	Aim	Findings	Subjects/Personality Instrument(s)
Busato et al. (2000)	To investigate the integration of intellectual ability, learning style (LS), personality, and achievement motivation, as predictors of academic success in higher education.	Intellectual ability and achievement motivation associated positively with academic success. LS appeared to be unrelated to students' success. For personality traits, Conscientiousness was positively associated with academic success (measured by study points and exam grades).	5FPT. 409 first year psychology students.
Duff et al. (2004)	To investigate relationship between approach to learning, personality factor, age, gender, prior educational, and AP.	Personality and learning approaches are poor predictors of AP. In terms of personality, the only trait that showed positive correlation with academic success was Conscientiousness (academic performance was measured by the GPA).	Cattell's 16PFi. 146 social science undergraduate students.
Pulford & Sohal (2006)	To investigate the influence of personality on higher education students' confidence in their academic abilities.	Results showed that personality did not influence the perception and confidence of higher education students. The level of Conscientiousness and Openness traits is shown to positively predict students' confidence level and reading and writing.	IPIP.
Komarraju et al. (2009)	To examine the relationship between personality, students' motivation and academic achievement.	Conscientiousness, Openness, Neuroticism, and Agreeableness are significant predictors of academic achievement (measured by self-reported GPA). The big-five traits are relatively important in explaining variance in students' GPA when compared with academic motivation.	NEO-FFI. 308 undergraduate students.
Chamorro-Premuzic & Furnham (2008)	To explore the potential of personality, intelligence and learning approaches in predicting AP.	Openness and Conscientiousness are shown to predict academic performance (measured by exam scores). Conscientiousness was found to be a better predictor of AP compared with intellectual ability.	NEO-PI-R. 158 undergraduate students.
Chamorro-Premuzic & Furnham (2003b)	To investigate the relationship between personality and AP at both primary and broad-level traits.	At the broad level, Conscientiousness is the most significant predictor of AP (measured by exam marks). Neuroticism and Extraversion are negatively correlated with AP, whereas Extraversion correlates only partially. At the primary level, dutifulness and achievement striving facets of Conscientiousness correlated significantly with AP; anxiety and impulsiveness facets of Neuroticism and gregariousness and activity facets of Extraversion are negatively correlated with AP.	NEO-PI-R. 247 undergraduate students.
Furnham et al. (2003)	To study relationships between personality, cognitive ability and belief about intelligence in predicting AP.	Conscientiousness is correlated positively with AP whereas Extraversion correlated negatively. Neuroticism, Openness and Agreeableness are not significant predictors of AP. Cognitive ability and belief about intelligence also did not significantly predict AP.	NEO-PI-R. 93 undergraduate students.

Chamorro-Premuzic & Furnham (2003a)	To investigate the relationship between personality traits and AP in two longitudinal studies.	In Study 1, Conscientiousness is positively and significantly correlated with AP but negative association was found for Neuroticism. In Study 2, Psychoticism is found to be a significant predictor and negatively associated with AP. Results from both studies support the incremental validity of personality traits in predicting AP.	Study 1: NEO-FFi (N=70). Study 2: EPQ-R (N=75).
Dollinger & Orf (1991)	To investigate the ability of Conscientiousness and Openness to experience in predicting AP.	Conscientiousness and Openness to experience were found to be significant predictors of AP. None of the personality traits correlated with final exam scores which involve higher order critical thinking (i.e. analysis of concepts, methods and theories).	NEO-PI (118 undergraduate students enrolled in personality psychology course).
Lounsbury et al. (2003)	To investigate the ability of intelligence, personality traits and work drive as predictors of AP.	Conscientiousness and Openness are the two traits significantly related with AP (measured by course grade). Regression analysis suggests that work drive added significant variance to the prediction of course grade beyond the personality traits.	Personal Style Inventory (PSI). 175 students in a senior-level psychology course.
Farsides & Woodfield (2003)	To investigate the potency of the big five traits in predicting academic success among higher education students.	Openness to experience and Agreeableness were positively correlated with academic success. Other traits (Extraversion, Conscientiousness, and Neuroticism) have no significant relationship with academic success.	NEO-FFI. 432 undergraduate students.
Nguyen et al. (2005)	To replicate previous studies about the relationship between personality and AP, and to explore the role of gender as a potential moderator of the relationship.	Conscientiousness positively and significantly predicted AP (measured through GPA and final course grade). Emotional stability positively predicted the students' performance among male students. Gender was consistently found to moderate the relationship between personality and AP.	IPIP. 368 undergraduate students enrolled in business courses.
Fruyt & Mervielde (1996)	To predict educational success based on two personality models: RIASEC and FFM.	FFM is shown to be able to explain variance in AP when compared with RIASEC. Of the five traits in FFM, Conscientiousness is significantly positively related with AP (measured by exam grades).	RIASEC and FFM (NEO-PI-R).
Conard (2006)	To investigate the incremental validity of the Big Five model in predicting AP.	Conscientiousness is shown to predict three academic outcomes (GPA, course performance and attendance) mediated by behaviour (attendance). The other traits did not provide incremental validity for AP. Results suggest that personality measurement has potential to be applied for college admission.	NEO-FFI (300 undergraduate students from Psychology classes) study spans for 3 years.
Poropat (2009)	The study performed a meta-analysis (MA) of the FFM and AP.	Of the five traits, Conscientiousness showed the strongest correlation with AP. Agreeableness and Openness are also found related to AP. These relationships are largely independent of intelligence and moderated by factors such as academic level (primary, secondary, or tertiary), age, and the interaction between academic level and age.	Cumulative sample size is approximately 70,000.

Our literature review regarding team personality composition indicates that most studies in the computing domain (i.e. IT/IS/SE) support diversity or a heterogeneous personality type in order to improve team performance (Karn & Cowling, 2006; Pieterse & Kourie, 2006; Bradley & Hebert, 1997). One of the main reasons highlighted was that heterogeneity helps in achieving greater performance due to "*the combined efforts of a variety of mental processes, outlooks and values*" (Karn & Cowling, 2006, p. 240). However, some literature in the psychology domain suggests that teams consisting of homogeneous personality are essential for team performance (e.g. Peeters et al., 2006; Kichuk & Wiesner, 1997). For instance, a team consisting of conscientious members is reported to perform better when compared with a team consisting of members with a heterogeneous level of Conscientiousness (Peeters et al., 2006). In this review we found no evidence in using FFM for studies regarding team personality composition in the computing related domain (e.g. IS/CS/SE) as most studies utilized MBTI in measuring personality.

In the studies that measure academic performance based on the FFM, there is a positive relationship between Conscientiousness and academic performance (Busato et al., 2000; Pulford & Sohal, 2006; Lounsbury et al., 2003). Nonetheless, most research focused on the association or correlation between academic success and personality traits, and therefore there is a lack of evidence on causal-effect type studies in the personality-related literature. This concern has also been raised by Boekaerts (1996) who mentioned that one major problem in understanding the effects of personality on students' learning is due to the lack of causal-effect type studies. Boekaerts suggests that researchers should not merely embark on conducting simple, correlational studies, but instead should study the effects of personality traits on various outcome variables such as achievement and learning strategies.

3.6 Summary

In summary, the literature reports various personality models for explaining individual differences including MBTI, FFM, Cattell's 16PF, and Eysenck. Of these measures, the FFM was chosen for this research due to its comprehensive nature in capturing human personality attributes. Besides, FFM is widely acceptable among personality psychologists and the validity and reliability of this model is also commonly reported in the literature. Our review of personality research from educational and team organizational perspectives suggests that there is an apparent relationship between personality and academic performance and team effectiveness. Knowledge of these areas is essential for development of hypotheses presented in the next chapter.

Chapter 4

OVERVIEW OF THE RESEARCH

In this chapter an overview of the research methodology and the set of formal experiments carried out as part of this research are presented herein. The explanation regarding experimentation in Software Engineering (SE) and the phases involved in conducting a typical formal experiment are also included. The research objectives are outlined using the Goal/Question/Metric framework followed by the formulation of hypotheses based on the evidence presented in the previous chapter. The research design employed in all formal experiments is also detailed. Finally, the set of instruments and materials used in the experiments, and the experimental and analysis procedures are also described.

4.1 Overview of the Research Process and Experimentation

The process followed in this research is shown in Figure 4.1. The planning for the series of formal experiments took place based on the research gaps discovered in the SLR. During the initial stage of the research, we identified the research objectives, the experimental design, the procedures, and the instruments to be used. In order to fulfill the human subject ethics requirements of the University of Auckland, we were required to seek approval from the University of Auckland's Human Participants Ethics Committee prior to executing each of the experiments conducted.

Experimentation in Software Engineering (SE) is usually a complex undertaking due to a greater number of factors or variables that should be considered in order to understand a specific phenomenon in software development (Juristo & Moreno, 2001). Nevertheless, experimental validation is a crucial part of the SE research domain in order to discover and test rigorous scientific knowledge of SE models, processes, methods, tools and techniques used in software construction (Juristo & Moreno, 2001). The research we carried out was a set of formal experiments under which a certain specific variable was controlled (i.e. the personality trait). Although it can be considered as a "controlled experiment", there are other variables which were not specifically controlled such as the students' gender, ethnicity, skills or abilities. The limitation in being able to control these variables was mainly due to the sample size employed in this research. These limitations are discussed under the threats to the validity of the results obtained from the set of experiments (see Chapter 11).

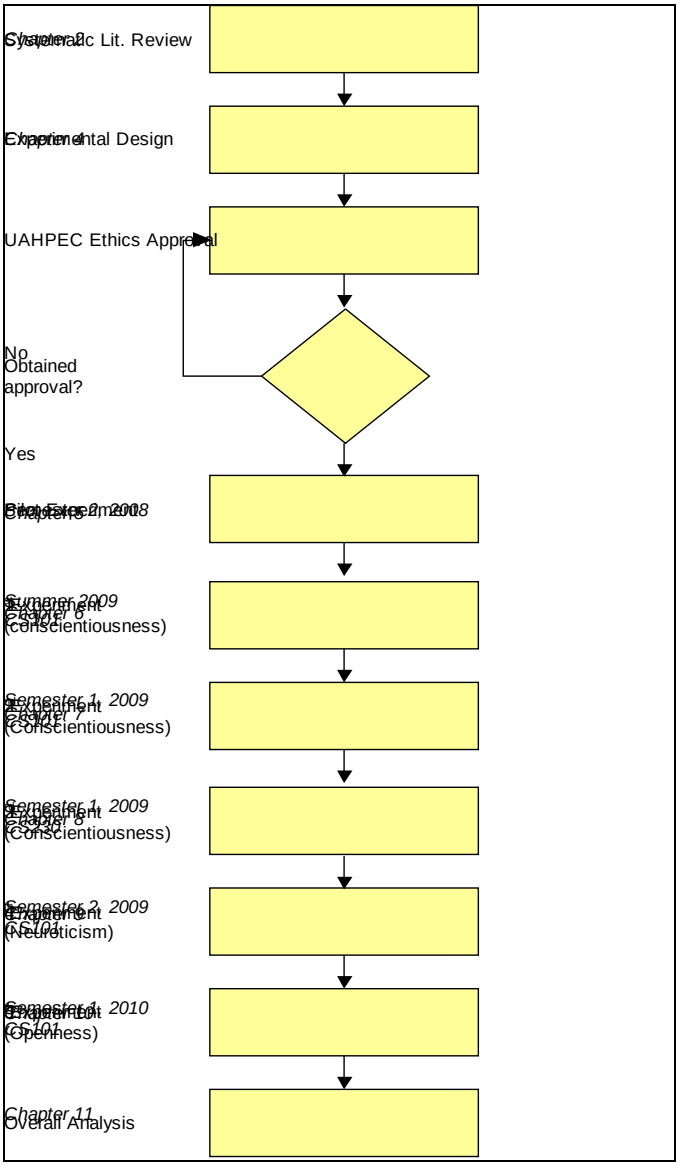


Figure 4.1 The Research process

According to Juristo and Moreno (2001), experimentation can be subdivided into four sequential of activities or phases as shown in Figure 4.2. The first phase involves the definition of the experiment's objectives. In this phase, the research objectives and the specific hypotheses are defined. During the design phase, the variables (both dependent and independent) are defined, the subjects or participants, and the type of research design employed in the study are also clearly identified. The execution stage takes places once the research design is completed. During this stage, the experiment is run according to the plan where data needs to be gathered and later analyzed. In the analysis phase, data analysis is performed according to the statistical procedure defined earlier during the design phase. During this phase, relationships between variables are analyzed and the hypotheses are tested. Inference statements can be made based on whether there is any statistical

significance observed from the analysis (i.e. to detect whether variation does exist among the group of observations due to the controlled variable).

The series of experiments (see Figure 4.1) were carried out according to the aforementioned phases. Every experiment employed a different set of independent variables and hypotheses, but shared similar experimental procedure and instruments (details of the variables used in this research are described in Section 4.5). This allows for data aggregation and overall analysis of findings during the final stage of the research.

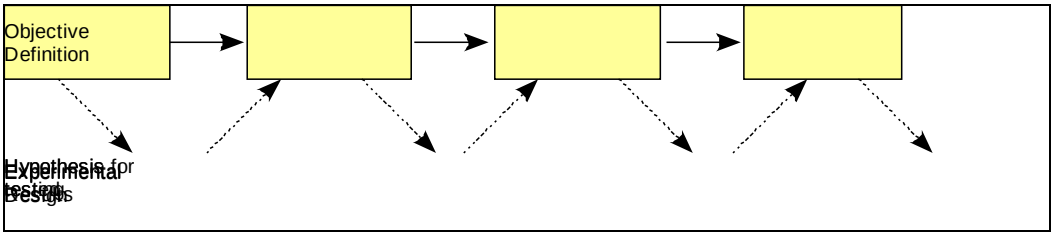


Figure 4.2 Process of experimentation in SE (Juristo & Moreno, 2001, p. 49)

Ethics approval was granted for three years starting from semester 2, 2008 (see Appendix B.1). In order to assess the suitability and clarity of the instruments (e.g. questionnaires), we conducted a pilot study prior to performing the actual experiment and data collection. A series of formal experiments were planned and carried out for four consecutive academic semesters starting from the 2009 Summer School (see Figure 4.1). During semester 1, 2009, there were two experiments executed in parallel involving two different types of courses taken by different levels of students. The aim of these twin experiments was to observe if there were any difference in effects on performance when similar personality traits or factors were used in each experiment. The findings from all experiments are aggregated and discussed in Chapter 11.

Each of the experiments investigates a different set of personality traits. Of the five major traits from the Five-Factor Model (FFM), three important traits, identified to be relevant for the research, were chosen: *Conscientiousness*, *Neuroticism*, and *Openness to experience*. The rationale and motivation for selecting these traits is detailed in Section 4.4. Every experiment lasts for a full academic semester involving undergraduate students attending weekly tutorial sessions monitored by a tutor, and assisted by several teaching assistants.

4.2 Research Context

Based on the evidence from our SLR, we found that personality was the most common factor investigated in previous PP studies (Salleh, Mendes, & Grundy, 2010). However, in terms of the effect or influence of personality towards PP's effectiveness, existing results were inconsistent. Research evidence also suggests that developers' personality is one of PP's most critical success factors (Cockburn, 2002). Therefore, the aim of this study was to improve the implementation of the PP practice as a pedagogical tool by focusing on

personality traits and investigate whether there would be any learning and/or performance improvements relating to pairing based on personality factors.

We have investigated a set of hypotheses in a series of PP experiments conducted at the University of Auckland between the periods of 2009-2010 involving two undergraduate courses: Principles of Programming (COMPSCI 101) and Software Design and Construction (COMPSCI 230). COMPSCI 101 is an introductory course for first-year students learning an object-oriented programming language, Java. During this course, students learn about basic programming concepts and create a few small applications as part of their assignments.

COMPSCI 230 is a more advanced course attended by second-year Computer Science students. The course consists of four major parts including software design using UML, object-orientation, database modelling, and JDBC programming. As part of their assignments, students are required to develop an object-oriented software and concurrent programming applications using a database. The following sub-sections describe the research objectives, and detail the hypotheses and the experimental design employed in this research.

4.3 Research Objectives

Our research objectives are outlined using the Goal/Question/Metric (GQM) framework (Basili, Shull, & Lanubile, 1999). The concept of GQM was developed by Basili and Rombach (1988) to represent a systematic approach for specifying a study's organizational framework. The GQM goal template contains five parameters that can be used to define a study's purpose (Basili et al., 1999). The GQM definition is shown in Table 4.1, and the purpose of all the experiments carried out as part of this research is outlined as follows:

Object of study: PP technique.

Purpose: To improve the effectiveness of PP as a pedagogical tool in higher education institutions.

Focus: To investigate the influence of personality as a psychosocial factor that can potentially affect the effectiveness of the PP practice in Computer Science/Software Engineering (CS/SE) courses/tasks.

Perspective: From the point of view of the researcher.

Context: In the context of undergraduate CS/SE students.

Table 4.1 GQM definition

Goal(s)	Question(s)	Metric(s)
To investigate the effect of personality differences towards successful pair configuration.	Do differences in personality traits affect PP's effectiveness?	Students' academic achievement measured by assignments, midterm test, and final exam scores.
To investigate the level of satisfaction of paired students.	Did students feel satisfied working in pairs?	PP questionnaire on satisfaction level.
To investigate the level of confidence of paired students.	Did students feel confident working in pairs?	PP questionnaire on confidence level.

4.4 Formulation of the Hypotheses

Existing literature suggests that the diversity or heterogeneity of personality among team members is a strong predictor of team success (Karn & Cowling, 2006; Pieterse & Kourie, 2006; Bradley & Hebert, 1997). In a follow up study of the effect of personality on the performance of SE teams, Karn and Cowling (2006) report that a team consisting of members with heterogeneous personalities worked well together. Similar findings were also reported by Busato et al. (2000), and Pieterse and Kourie (2006). Their studies however, were conducted in the context of teams consisting of four to five members.

The effects of personality were also investigated in PP studies involving peer or pair collaboration. Choi (2004) reports that PP works effectively for paired students with different personality types. Another study by Sfetsos et al. (2009) also reports that pairs consisting of heterogeneous personalities performed better than pairs with the same personality type. Most existing PP studies measured personality type using the Myer-Briggs Type Indicator (MBTI) as a personality assessment method (see Chapter 2). Although MBTI was found to be very popular and widely used in research in the computing and business domains, there is evidence that the Five-Factor personality dimensions are a robust taxonomy of personality (Barrick et al. 1998; Burch & Anderson, 2008). This evidence has motivated us to employ the Five-Factor Model (FFM) in this research.

As far as we are aware, there are no available theories that link the Five-Factor Model (FFM) with PP. Hannay et al. (2010) also shared a similar view when they mentioned that "*there were no explicit references to theory for explaining effects of personality on pair programming*" (p. 65). Nevertheless, personality traits were predicted to play an important role in undertaking programming tasks, as asserted by Weinberg (1971). Therefore, in order to investigate whether FFM's personality traits have significant influence on PP, differences in personality can be operationalized by forming pairs consisting of students with different levels of personality traits. In order to investigate the effect of personality differences on PP's effectiveness, the following general null hypothesis was proposed:

> *H_O: Differences in personality traits do not affect the effectiveness of students who pair programmed.*

which is contrasted by the following alternative hypothesis:

> *H_A: Differences in personality traits affect the effectiveness of students who pair programmed.*

In this research, of the five broad traits proposed in the FFM, the three major traits reported to be important educationally and relevant for higher education in previous studies are: Conscientiousness, Openness to experience and Neuroticism (De Raad & Schouwenburg, 1996). Previous findings showed Conscientiousness to consistently positively predict educational success (Busato et al., 2000; Duff et al., 2004; Chamorro-Premuzic & Furnham, 2003b). High Conscientiousness is always related to being a high achiever, organized, and thorough, whereas low Conscientiousness possesses the opposite traits such

as a low need for achievement, being unprepared and disorganized (McCrae & John, 1992). Thus, in the present research this factor is believed to affect PP's effectiveness. We hypothesize that pairs consisting of highly conscientious students are expected to achieve better academic performance than pairs presenting low levels of Conscientiousness. Hence, in order to investigate the above hypotheses, more specific hypotheses were developed:

Hypotheses 1:

> *H1_O: Differences in Conscientiousness level do not affect the effectiveness of students who pair programmed.*
> which is contrasted by the following alternative hypothesis:
> *H1_A: Differences in Conscientiousness level affect the effectiveness of students who pair programmed.*

Of the five personality constructs, Neuroticism (or lack of emotional stability) is the factor that deals with feelings of anxiety, self-consciousness, impulsiveness, and vulnerability (De Raad & Schouwenburg, 1996; Costa & McCrae, 1995). Evidence suggests that Neuroticism is negatively correlated with academic performance due to the effects that traits such as anxiety and impulsiveness have (Chamorro-Premuzic & Furnham, 2003b). It should however be noted that there is some evidence from organizational psychology that in certain conditions anxiety and Neuroticism may actually facilitate performance (Burch & Anderson, 2008). On a positive side, emotional stability is consistently related to self-efficacy, which in turn, affects performance (Schmitt, 2008; Barrick et al., 1998). Barick et al. (1998) report that teams comprising more emotionally stable members (i.e. low Neuroticism) are likely to achieve higher performance when compared with teams that include even a single member who is emotionally unstable. Therefore, we posited that the level of Neuroticism may influence the academic performance of students practicing PP; hence the following hypothesis was proposed:

Hypotheses 2:

> *H3_O: Differences in levels of Neuroticism do not affect the academic performance of students who pair programmed.*
> which is contrasted by the following alternative hypothesis:
> *H3_A: Differences in levels of Neuroticism affect the academic performance of students who pair programmed.*

In terms of Openness to experience, literature in the personality psychology reports that Openness to experience facilitates the use of learning strategies, and that students with a relatively high level of Openness are described as being foresighted, intelligent, and resourceful (De Raad & Schouwenburg, 1996). Farsides and Woodfield (2003) report that

there is a positive correlation between academic success and Openness to experience among undergraduate students. Therefore, this trait may play a role in the effectiveness of students who pair programmed; hence the following specific hypothesis was proposed:

Hypotheses 3:

> *H2_O: Differences in levels of Openness to experience do not affect the academic performance of students who pair programmed.*
> which is contrasted by the following alternative hypothesis:
> *H2_A: Differences in levels Openness to experience affect the academic performance of students who pair programmed.*

The selection of personality traits as variables may provide an advantage in overcoming the problem of bad pairing experiences reported in some PP studies (Layman, 2006; Ho, 2004). The discomfort or incompatibility working with a partner might be due to a mismatch in psychosocial aspects such as personality and gender combinations, or in competency aspects, such as skill or experience levels. The findings from our SLR indicate that students prefer to work with a partner who is at similar skill level as theirs (Salleh et al., 2010). Cockburn and Williams (2001) highlight that understanding the social aspects of PP is critical for attaining the success of the practice. This is mainly because the PP practice is a collaborative process involving interaction and communication between two people working together to achieve a common set of goals. As different people possess different behavior and opinion, understanding how the two students can work best together is imperative to the success of PP as a pedagogical tool.

4.5 The Research Design

Schloss and Smith (1999) defined research design as "*an objective and complete description of the methodology employed by the researcher*" (p. 53). The choice of research design depends on the research questions being asked or hypothesis being investigated, and the sample being studied (Schloss & Smith, 1999). This is used to increase the credibility of the research results and to enable conclusions to be interpreted with confidence. Creswell (2003) reports the two major strategies for quantitative inquiry:

i) Experiments

The three types of experimental research are true experiments, quasi-experiments and correlational studies (Creswell, 2003). The difference between true experiments and quasi-experiments lies in the approach to assign experimental subjects to groups of observation. In true experiments, subjects are assigned to a treatment randomly as opposed to in a quasi-experiment. On the other hand, a correlational study seeks to investigate the association or relationship between two or more variables (a detailed definition of the terminology is given below).

ii) Surveys

A survey provides a way to quantify trends, attitudes, or opinions of a population based on studying some characteristic or behaviour of the population using questionnaires or structured interviews as a means of gathering data (Creswell, 2003). Examples of survey design are cross-sectional and longitudinal studies (Creswell, 2003). A cross-sectional study represents a study where data are collected at "one point in time", whereas in longitudinal studies, data are collected over a certain period of time (Creswell, 2003).

In the present study, the two types of empirical investigations employed were true experiments and correlational studies. Another terminology used by the SE community when referring to a true experiment is "*formal experiment"* or "*controlled experiment"* (Sjoberg et al., 2005; Pfleeger, 1995). In this thesis, we use the term "formal experiment" as that is the term widely used in SE empirical research (Pfleeger, 1995). A formal experiment is important in any empirical SE research because it enables a researcher to understand the causes of the phenomena being studied (Juristo & Moreno, 2001). Formal experiments are typically performed under a systematic and controlled environment, thus allowing a researcher to draw conclusions regarding cause-and-effect relationships (Ogot & Okudan, 2006; Pfleeger, 1995). Such a control over variables makes the results of formal experiments generally applicable to a much wider population, in contrast to those of case studies or surveys (Pfleeger, 1995).

Pfleeger (1995) mentioned that some of the significant benefits of conducting experimental studies in SE are: (i) to confirm theories or claims about the best approaches to be used among the many proposed SE techniques, tools and methods; (ii) to explore relationships among various characteristics of software products and resources used in software development because understanding these relationships is pertinent for a project's success; (iii) to evaluate the accuracy of SE models in predicting the outcomes of a project; (iv) to validate software measures i.e. whether the measures are truly reliable in capturing the value of a specific software attribute.

Before discussing the detailed aspects of the research design employed in this research, it is important to understand the terminology used to describe the research design. Thus, a brief definition is given below:

- *Independent variable (IV)*

  An IV is described as a variable that is directly used and manipulated by a researcher, which could possibly cause, influence, or affect another variable (Leedy & Ormrod, 2005; Creswell, 2003). This variable is also known as a "manipulated", "factor" or "predictor" variable because it is used to predict the result of the experiment (Juristo & Moreno, 2001; Creswell, 2003). Examples of IVs are software testing techniques (i.e. code reading, functional testing etc.), programming approaches (i.e. Aspect J, Java etc.), personality (i.e. low Conscientiousness, high Conscientiousness, etc).

- *Dependant variable (DV)*

  A DV is a variable that is "*potentially influenced by the independent variable*" (Leedy & Ormrod, 2005, p. 218). It can also be referred to as the outcome of an experiment which can be measured in order to examine the effects of the IV (Creswell, 2003). This variable is also known as the "response variable" which is "*expected to change or differ as a result of applying the treatment*" (Fenton & Pfleeger, 2001, p. 128). Examples of DVs are program size (such as measured by number of classes, methods etc.), process effort (measured as person-hours), cost (measured as dollar per month), and academic performance (measured as final grades, test scores etc.).

- *Treatments*

  Fenton & Pfleeger (2001) define treatment as "*the new method or tool you wish to evaluate (compared with an existing or different method or tool)*" (p. 127). The purpose of any formal experiment is to determine whether the treatment is beneficial with regard to the outcomes we are measuring (Fenton & Pfleeger, 2001). It is also known as the "levels" or "alternative" of an investigated factor (Juristo & Moreno, 2001). For example, the treatments for investigating the effects of a pedagogical technique (IV) on learning programming would be pair programming and solo programming.

- *Experimental Unit*

  The experimental unit refers to the object of an experiment where the treatment is being applied (Fenton & Pfleeger, 2001). For example, in a study investigating the effectiveness of pair programming as a pedagogical approach in learning programming in comparison to solo programming, the experimental unit would be the students using PP or working solo in the lab or classroom.

- *Experimental subjects (participants)*

  The experimental subjects are the people who directly participate in the formal experiment (Juristo & Moreno, 2001). Examples are a team of developers in software companies, or undergraduate students enrolled in CS courses.

- *Intervening or mediating variable*

  The intervening variables are variables that intervene or mediate the effects of the IV on the DV (Creswell, 2003). Thus, it "stands between" the independent and dependent variable and it helps explain the reason why the IV affects the DV. For example, in investigating the effects of "teaching method" (IV) on the students' academic performance (DV), intervening variables that may influence students' performance are possibly "intelligence" and "study skills".

- *Moderating variable*

  The moderating variables are variables that have an effect on the relationship between the IV and the DV. They represent the third variable that may transform the original relationship between the IV and the DV. A moderating relationship occurs when the relationship between two variables (IV and the DV) depends on the level of

moderating variables (Baron & Kenny, 1996). For example, a relationship between programmers' collaboration and team effectiveness may be moderated by amount of communication that occurs between team members.

- *Experimental error or confounding error*
  These errors are due to extraneous factors that may possibly *"influence the characteristics under study but which have not been singled out for attention in the investigations*" (Pfleeger, 1995, p. 231). For instance, subjects may disturbed by loud noises from another room, or distracted by the room temperature. A good experimental design should aim to minimize the effects of the confounding errors (Pfleeger, 1995).

Pfleeger (1995) discussed the three major principles in addressing the issues of confounding errors. These are *replication*, *randomization*, and *local control*. Replication allows for an experiment to be repeated and hence for the effects of confounding errors to be identified and increase confidence in the results obtained. Randomization helps reduce bias caused by any uncontrolled sources of variation (Pfleeger, 1995). Local control refers to the degree of control over the assignments of subjects in the experiment. This could be achieved by two design types: *blocking* and *balancing*. Blocking helps in organizing the experimental unit by allocating them into a homogeneous group. Balancing allows an equal number of subjects be assigned into a treatment (Pfleeger, 1995).

The visual model of research design used in our study is illustrated in Figure 4.3. This research model was derived based on the initial framework for research on PP proposed by Gallis, Arisholm & Dyba (2003). The model shows the interaction between the variables and the expected or observed outcomes (i.e. in terms of how the treatment would benefit the experimental subjects). Although three important personality traits were investigated throughout a series of experiments, each experiment focused on only a single personality trait (e.g. Conscientiousness). Thus, personality trait was a "factor" used to predict the performance of paired students. Based on the personality scores, the personality trait can be classified into three levels: low, medium, and high. Therefore participants were allocated into pairs according to their personality level. For instance, pair configuration for the personality trait Conscientiousness was designed as below:

Pair (C_{High}, C_{High}) → denotes a pair combination where both students have high levels of Conscientiousness.

Pair (C_{Low}, C_{Low}) → denotes a pair combination where both students have low levels of Conscientiousness.

Pair (C_{Medium}, C_{Medium}) → denotes a pair combination where both students have medium levels of Conscientiousness.

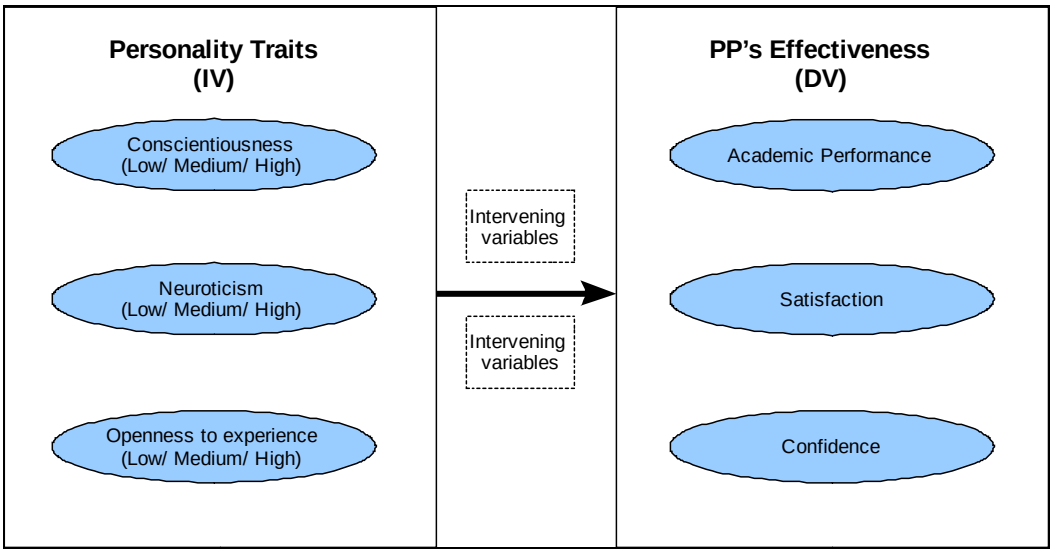


Figure 4.3 Visual model of research design

Based on this research model, the research design employed in this study was a "single-factor between-group design" (Morgan, Leech, Gloeckner, & Barrett, 2004; Pfleeger, 1995). The "between-group" design was used because each student or participant in the research was assigned into only one condition or group for every treatment (Morgan et al., 2004). The treatment here refers to the pairing allocation based on the participants' personality trait levels. Thus, each participant can be assigned into only one of the three groups mentioned above (i.e. low, medium, or high). The only exception to this design was the first experiment (see Chapter 6) where the groups used related to the combinations homogeneous vs heterogeneous personality trait. In this particular experiment, the homogeneous group consists of paired students with similar personality and the heterogeneous group represented paired students of mixed personality. The former group was known as the "control group" and the latter as the "experimental group".

PP's effectiveness was the outcome to be measured on every experiment. According to our SLR, measuring PP's effectiveness could be achieved using "academic performance", "technical productivity", "program quality", or "satisfaction" (Salleh et al., 2010). Since our research aimed at facilitating CS/SE students' learning through the practice of PP, the metrics we chose to use to measure PP's effectiveness were "academic performance", students' "satisfaction" and students' "confidence". Hence, PP's effectiveness was the dependent variable, and personality trait the independent variable.

In this research, the tutorials' topic varied from week to week. Therefore, the experiments were designed in such a way to minimize the confounding error which might occur due to differences in the tasks' complexity assigned to the students. Hence, the tasks and exercises assigned to students remained the same throughout a week. In this regard, the blocking variable applied to all the experiments was the topic for exercises. Table 4.2 summarizes the attributes used in this research defined according to the research design terminology.

Table 4.2 Study attributes and metrics

No.	Term	Attribute(s)	Metric(s)
1.	Independent Variable	Personality trait (Conscientiousness, Neuroticism, and Openness).	Personality trait scores.
2.	Dependent Variable	PP's effectiveness.	Academic performance, satisfaction and confidence.
3.	Treatment	Pair allocation by personality levels (low, medium, high).	-
4.	Blocking variable	Tutorial's topic.	-
5.	Experimental Subjects	Undergraduate students enrolled in the course selected for the experiment.	-
6.	Experimental Unit	Students pairing for the tutorial in a particular week.	-
7.	Experimental Error	Level of complexity of tasks.	-

In all the experiments, academic performance was measured using assignments, a midterm test, and final exam scores. The levels of satisfaction and confidence were measured using a questionnaire where all questions employed a five-point Likert-scale. The four common types of measurement scales applied in SE and social sciences research are: (i) Nominal, (ii) Ordinal, (iii) Interval, and (iv) Ratio scale (Leedy & Ormrod, 2005; Juristo & Moreno, 2001). In this research, students' academic performance in assignments, midterm test and final exam were measured based on a ratio scale, whereas satisfaction and confidence levels were measured based on an ordinal scale. The personality scores were measured based on interval scales because the scores were represented on a numerical scale (between 0 and 100) and there are no "true" zero scores (i.e. the scale does not represent the absence of certain personality trait being measured). Table 4.3 lists the measurement scale types employed in this research.

Table 4.3 Attributes and measurement type scales

No.	Attribute(s)	Measure
1.	Academic performance (assignments, midterm test and final exam)	Ratio Scale
2.	Personality scores (Conscientiousness, Extraversion, Agreeableness, Neuroticism, and Openness to experience)	Interval Scale
3.	Personality level (low, medium, high)	Ordinal Scale
4.	Date of Birth	Interval Scale
5.	Gender	Nominal
6.	Ethnicity	Nominal
7.	Work experience related to computing (number of years)	Ratio Scale
8.	Programming competency	Ordinal Scale
9.	Experience in PP	Nominal
10.	English as the first language	Nominal
11.	Satisfaction level	Ordinal Scale
12.	Confidence level	Ordinal Scale

4.6 Instrumentation and Materials

There were six types of instrumentations and materials used in the experiments:

(i) Participant Information Sheet (PIS).
(ii) Consent Form.
(iii) Personality Test (IPIP-NEO).
(iv) Demographic Survey Form.
(v) PP Questionnaire.
(vi) Pair Allocation program.

The PIS described the nature of the experiment by highlighting its major purpose and the activities involved, thus the PIS provides sufficient information to the participants for making a reasonable judgment on whether to participate in the experiment (see Appendix B.2 for the PIS used in our research). Participation in this research is on a voluntary basis; therefore subjects are given the right to withdraw from the study at any time before the end of semester. Participants who are willing to participate are given a consent form (See Appendix B.3). A consent form lists the statements indicating the nature or conditions of participation and allows participants to indicate their agreement to participate by signing the form.

The short version of the *International Personality Item Pool Representation of NEO PI-R™* (IPIP-NEO) was employed in order to measure the personality traits of participants[2] (see Appendix D). The IPIP-NEO was developed based on the *International Personality Item Pool* (IPIP), a scientific collaboratory for the development of personality measurement scale and other individual differences (Goldberg, 1999). The original version of IPIP-NEO contains 300 items whereas the short version contains 120 items of which the descriptions were authored by Johnson (2008). Although the original version of IPIP-NEO provided a more reliable result, the short version was reported to measure exactly the same traits and to also present acceptable measurement reliability (Johnson, 2008). The selection of IPIP-NEO as the personality test used in this research was due to two major reasons: i) It was developed based on the FFM framework, and ii) It provides a Web interface for collecting and scoring calculation of personality responses, which is more efficient compared with the paper-based version (Buchanan, Johnson & Goldberg, 2005; Gosling, Vazire, Srivastava, & John, 2004).

Each item in the IPIP-NEO personality inventory was indicated based on the five-point Likert-scale ranging from “Very Inaccurate” to “Very Accurate”. The test produces scores in a numerical scale; with 0 and 99 representing the lowest and the highest scores for each trait, respectively. These numerical scores were then “translated” into an ordinal scale (i.e. low, medium, high) for the purpose to assign pairs. Based on the suggestion described by Johnson (2008) the personality traits were classified into low, medium or high level based on the range of scores shown in Table 4.4.

[2] The personality test is available at the public domain http://www.personal.psu.edu/ /j5j/IPIP/

Table 4.4 Personality scores level

Scores	Lowest 30%	Middle 40%	Highest 30%
Level	Low	Medium	High

In addition to the online personality test, experimental subjects were administered with a pre-test questionnaire to gather their demographic information, work experience, and to rate their programming competency level (see Appendix B.5). The questionnaires were distributed to the students during the first course lecture session where an introduction to the formal experiments was given. The PP questionnaire was used to gather participants' feedback regarding their experience working in a pair, and also to measure participants' satisfaction and confidence level working with their partner (see Appendix B.4). The feedback was rated using an ordinal five-point scale ranging from "Strongly Disagree" to "Strongly Agree". The satisfaction level was rated according to an ordinal scale ranging from "Very Dissatisfied" to "Very Satisfied", whereas confidence level was rated from an ordinal scale ranging from "Very Low" to "Very High".

Finally, *Pair Allocation* (PALLOC) software was used in order to automate the process of pairing formation. It is a Java-based application that connects to a MySQL 5.1 database server and runs under the Eclipse 3.2 environment. Based on the weekly list of students provided by the tutor, PALLOC generates the pairing list in Microsoft Excel's document format. The database structure and the overview of design model used in PALLOC are available in Appendix C.1 and Appendix C.2.

4.7 Experimental Procedure

For each of the experiments carried out, the experimental procedure was as follows: At the start of the academic semester, we would approach the experimental subjects in the first course lecture. During that lecture, students were given an overview of the experiment, including a brief explanation on PP and at the same time the PIS, the demographic survey form and the consent form were distributed for signing. In addition, participants' personality profiles were also gathered during the first two weeks of the semester using the online IPIP-NEO test. The results of the personality profiling were then employed to allocate partners. For this purpose, the score of a specific personality trait (e.g. Conscientiousness, Openness or Neuroticism) was used as a basis to generate the pairing list randomly within each group (i.e. low, medium or high). This process was executed weekly by using the PALLOC program.

We have carried out one experiment involving COMPSCI 230, whereas the other experiments were carried out involving COMPSCI 101. The nature of the courses is given below:

(i) COMPSCI 101 (Principles of Programming)

This course is designed to be enrolled in by first-year undergraduate students. The course consists of ten weeks of lectures and weekly compulsory tutorials. Programming concepts and theories were explained during formal lecture hours, and students were

given preparation sheets to be completed before attending a tutorial. Our experiment took place during the compulsory closed weekly labs or weekly tutorial sessions run by a tutor and a few teaching assistants (TA). During a tutorial session or closed lab, students were required to submit the preparation sheet to the TAs to be graded and they were also required to solve a minimum of two programming problems with their allocated partner. Every tutorial lasted for two hours where the first 45 minutes were used by the tutor to explain the topic, and the remaining 75 minutes were allocated for students to solve the exercises in pairs. To allow for "pair jelling" (Williams, Kessler et al., 2000), students worked with their partner for an initial period of 30 minutes. They were then required to swap their roles. The swapping process was instructed by the tutor to ensure that every student had experience fulfilling both roles i.e. taking turns at being the driver and the navigator. The exercises given during the tutorials were graded, thus contributing towards their final grade. In addition, assignments, a midterm test and final exam were also graded, however completed individually. Students' grades in this course were determined by the scores on the tutorial exercises, assignments, a midterm test, and a final exam.

(ii) COMPSCI 230 (Software Design and Construction)

This course is designed to be enrolled in by second-year CS students. The course consists of ten weeks of lectures and weekly non-compulsory tutorial. In this course, tutorials were prepared for students needing help in understanding the subject matter; hence attendance is not mandatory. Students intended to attend a tutorial were requested to inform the tutor prior to the session. This is to enable us to assign students into pairs. During the tutorial, paired students were given a set of exercises by the tutor. These exercises were not graded but were discussed at the end of the tutorial. The duration of tutorial is only one hour. Students' grades in this course were determined based on the assignments scores, a midterm test, and a final exam scores.

In both courses, during a tutorial session or closed lab, students worked in pairs with their allocated partner. Participants' feedback working with their partner was gathered for every session, thus each of the tutorial sessions was treated as an independent formal experiment. Before the end of each tutorial/lab students provided feedback about working with their partner by filling out a short questionnaire (see Appendix B.4). The experiments aim to measure the effect of pair personality composition towards the academic performance of the paired students. Thus, the same research design is used every week until the end of the semester.

4.8 Analysis Procedure

Analysis of the data involved the selection of inferential statistics and software packages used to carry on the analysis (Morgan et al., 2004). The inferential statistics were used to make

inferences or deductions regarding the population based on the data that had been collected and analyzed (Morgan et al., 2004). Morgan et al. (2004) described the two basic inferential statistics as the “difference inferential statistics” and “associational inferential statistics”. The *difference inferential statistics* draw a conclusion about the population by computing the mean differences between the investigated groups. Thus, it can be used to perform the statistical testing on the set of hypotheses used in a study. On the other hand, the *associational inferential statistics* make inferences about the relationship or association between the studied variables (Morgan et al., 2004).

In order to test the null hypotheses, we used a set of parametric statistical tests. This is due to the type of measurement scale used to measure the dependent variables (Pallant, 2007). The parametric tests appropriate for comparing mean differences in this research are One-Way *Analysis of Variance* (ANOVA) and *Multivariate Analysis of Variance* (MANOVA). A One-Way ANOVA is chosen when there is only one independent variable with three or more levels and at least one continuous dependent variable. In the present research, ANOVA was used for analyzing mean differences in academic performance between different personality groups (i.e. low, medium, high). In the case of MANOVA, analysis of the data can be done simultaneously based on a linear combination of several dependent variables that are correlated at a low to moderate level (Leech, Barret & Morgan, 2005). One advantage of using MANOVA is that it “controls” or adjusts for the likelihood of getting a Type 1 error or “inflated Type 1 error” (Pallant, 2007). A Type 1 error is the probability or likelihood of getting the null hypotheses incorrectly rejected (Cohen, 1988). However, there are several assumptions that have to be complied with before choosing MANOVA for analysis. The main assumptions for MANOVA are in regards to the minimum sample size required for analysis, the multivariate normality, and the homogeneity of variance/covariance matrices (Pallant, 2007; Leech et al., 2005). The multivariate normality means that the distribution of scores on the dependent variable is normally distributed. The homogeneity of variance refers to the variances for each dependent variable (i.e. variability of scores) which should be approximately equal in all groups (Pallant, 2007; Leech et al., 2005).

In order to measure whether the independent variable had an effect on the level of “satisfaction” and “confidence”, we applied a non-parametric test such as Mann-Whitney and Kruskal-Wallis. A non-parametric test is chosen when the assumptions for parametric test are violated, e.g. when the dependent variable’s data are non-normally distributed or data are measured on an ordinal scale (Morgan et al., 2004). The Mann-Whitney statistical test was used to compare the mean ranks for satisfaction and confidence level between the controlled and experimental groups. When the comparison involved three or more levels or groups, the Kruskal-Wallis test was used.

In terms of measuring the association or relationship between continuous variables (e.g. personality traits and academic performance), we used the bivariate Pearson correlation Finally, the statistical software package employed in this study was SPSS version 17.

4.9 Summary

The aim of this chapter is to put the research process into context. As such, the stages involved in the research process were explained, the formulations of hypotheses were described, and the experimental design, procedure and instruments were also detailed. The research consists of a series of formal experiments executed in several academic semesters involving introductory CS courses enrolled in by first-year and second-year undergraduate students.

These experiments aim to investigate the effects of personality traits on PP's effectiveness measured by students' academic performance. Personality traits, as measured using the FFM framework, consist of five broad traits; our research, however, focused on the effects of the three important traits relevant for tertiary education: Conscientiousness, Openness to experience, and Neuroticism. The following chapter describes the pilot experiment conducted at an early stage of research. This was used to provide an initial test and refinement of the experimental approach, the instruments and analysis techniques as described in this chapter. The remaining chapters describe each of the formal experiments by detailing the experimental set-up, the data captured, the statistical analysis of data, and discuss the results obtained based on the analysis.

Chapter 5

THE PILOT EXPERIMENT

This chapter describes an initial pilot experiment which was conducted at the University of Auckland during the Second Semester of 2008. The experimental subjects were year-two undergraduate students attending the Software Construction and Design (COMPSCI 230) course. The main purpose of the pilot experiment was to validate the instruments, and the research design used in our research. The objectives, context, and the investigated hypothesis are explained in the following sections. Finally, the results obtained are discussed and limitations of the procedures and instruments used in the study are identified. We used these results to inform the design and analysis of our subsequent experiments, described in the following chapters.

5.1 Pilot Experiment's Objectives

Our research aims to improve the effectiveness of pair programming (PP) as a pedagogical tool for Computer Science/Software Engineering (CS/SE) education by investigating the effects that personality differences among paired students may have on PP's effectiveness. Students' personality profile was measured based on the Five Factor personality model.

The primary purpose of our pilot experiment was to test the arrangement and feasibility of the instruments to be used in our set of follow-up experiments. One of the aims of the pilot study was to ensure that there would be no misunderstanding from the participants regarding the items used in the questionnaires and other instruments. Participants' comments and suggestions were taken into consideration so that the clarity and quality of the instruments used in the pilot were improved upon for subsequent experiments. Our pilot study was conducted in some of the tutorial sessions optionally attended by the students. The tutorials were held weekly and monitored by a tutor. Each tutorial session lasted for an hour.

5.2 Pilot Experiment's Context

A pilot experiment was executed in the weekly tutorial sessions of the Software Design and Construction (COMPSCI230) course during the Second Semester of 2008. Each of the tutorial sessions was treated as an independent experiment. This course consists of four major parts, including software design using UML, object-orientation, database modeling, and JDBC programming, where students were provided with the opportunity to apply PP to design-related tasks in addition to programming tasks. Students willing to participate in the pilot study were required to sign a consent form to fulfill the ethical requirements of the University of Auckland's Human Participant Ethics Committee (See Appendix B.3). Students

enrolled in this course have gained a basic understanding of programming from the Principles on Programming course (COMPSCI 101), taken during their first year of study.

5.3 Hypothesis

The hypothesis investigated in the pilot experiment was to seek out whether differences in students' personality profile when pairing had a major influence on the effectiveness of students' academic performance. Within the scope of the pilot study, differences in personality were operationalized by forming pairs consisting of students with different levels of *Conscientiousness*. The detailed derivation of the hypothesis was discussed in Chapter 4. We restate the null and corresponding alternative hypothesis as follows:

H_O: Differences in personality trait Conscientiousness do not affect the effectiveness of students who pair programmed.

H_A: Differences in personality trait Conscientiousness affect the effectiveness of students who pair programmed.

As discussed in Chapter 3, Conscientiousness was chosen because it has been shown to be the most consistent predictor of academic performance in the psychology literature (Busato et al., 2000; Duff et al., 2004; Furnham et al., 2003). Previous findings also showed that a PP team of heterogeneous personalities achieved better performance than that of similar personalities (Choi, 2004; Sfetsos et al., 2009). Therefore, the pilot experiment compared the performance between similar and mixed personality pairs. Table 5.1 shows the categorization of pairs according to personality differences on the Conscientiousness factor. Pair (C_{High}, C_{High}) denotes a pair consisting of students with very similar personality (higher scores on Conscientiousness). Meanwhile Pair ($C_{High,}$ C_{Low}) refers to a pair of students of very different personality (higher and lower scores on Conscientiousness). High Conscientiousness is represented by scores above 70, whereas scores below 30 represent low Conscientiousness and scores between 30 and 70 represent medium Conscientiousness.

Table 5.1 Personality differences

Similar Personality	Mixed Personality
Pair (C_{Low}, C_{Low})	Pair (C_{Low}, C_{Med})
Pair (C_{Med}, C_{Med})	Pair (C_{Med}, C_{High})
Pair (C_{High}, C_{High})	Pair (C_{Low}, C_{High})

The pilot study also looks into the association or correlation between students' personality scores and their academic achievement, level of satisfaction with their pair-programming experience and paired students' confidence level in solving the exercises.

5.4 Variables

During the pilot study, a series of "mini experiments" were conducted (one per tutorial) to investigate how variations in the personality trait Conscientiousness would affect PP's

effectiveness as a CS/SE pedagogical tool. Hence, Conscientiousness was the independent variable, and PP's effectiveness was the dependent variable. In measuring PP's effectiveness, students' academic performance such as marks in tutorial exercises, assignments, and midterm test were compared. Levels of satisfaction and confidence of paired students were also evaluated using questionnaires consisting of five-point Likert-scale items (see Appendix B.4).

5.5 Experimental Procedure

As discussed in Chapter 4, the projected hypotheses were investigated using a "single factor between-group design" as the experimental design. This design allows each participant in the study to experience only one condition or group. This means that, in a particular lab session, a student was only assigned either into a pair of similar or mixed personality group. Students' personality data was gathered using an online IPIP-NEO test[3] during the first two weeks of the semester. The personality profiling of students was used to allocate partners randomly within each group using the Conscientiousness factor scores (e.g. A student with high Conscientiousness score was paired with someone with a low Conscientiousness score to form a mixed personality pair).

During the tutorials, participants worked with their assigned partners and were given a set of exercises later graded by the principal researcher. The assessment of the tutorial exercises given on each tutorial aims to measure the amount of students' learning, and as such it is the dependent variable used to represent PP's effectiveness. Tutorial assessments did not count towards the final grades obtained by the students. In addition to assessing tutorial exercises, we also used as measures of PP's effectiveness assignments, test, and final exam scores. All coursework assessment, including assignments, tests, and exams, were assessed individually.

5.6 Preliminary Results

Since the course was taught by three different lecturers throughout the semester, the consensus was to collect data from a block of weeks that were taught by the same lecturer. Due to constraints imposed by some of the course's lecturers, data was gathered for only three weeks of tutorials (from the 8th week of the semester to the 10th week). These experiments were referred to respectively as *exp1*, *exp2* and *exp3*. Note that prior to running these three weeks of experiments, students worked individually during the tutorial sessions.

There were 130 students enrolled for the COMPSCI 230 course during the second semester of 2008. Of these, 31 students (26%) consented to participate and completed the personality test. Ninety-four percent (94%) of the subjects were male, 6% female (29 male, 2 female). We had collected data from 18 students (9 pairs) who had attended the tutorial sessions for both *exp1* and *exp2*, and only 4 students from *exp3*. Due to the very small

[3] The personality test can be accessed at: http://www.personal.psu.edu/j5j/IPIP/

sample size in *exp3*, only the data analysis from exp1 and exp2 will be presented. Note that attendance at the tutorials was not compulsory given that the tutorial sessions were only intended to help students in gaining better understanding of the topic learned during the lecture(s).

Of the 31 students who consented to participate in the study, most (77%, 24 students) indicated that they did not have any work experience. The programming competency was assessed by a survey asking the students to rate their competency on a scale from 1 (very poor) to 5 (outstanding). In terms of programming competency, of the 31 students, 17 (55%) rated their competency as fair, 5 (16%) rated their competency as good or outstanding, whereas 9 (29%) rated their competency as poor. Subjects' age in this pilot study ranged from 19 to 46 years (mod = 20 years).

Table 5.2 shows the minimum, maximum, standard deviation, mean and the median scores for each of the five personality factors for the 31 participants. The boxplot in Figure 5.1 shows the distribution of students' personality scores. The rectangle inside the boxplot shows the interquartile range, with the median represented by the line drawn across the box. The median value for the Neuroticism was the highest amongst the five personality factors, and the lowest median was found in the Openness to experience. The black dot outside the distribution range was an outlier, and in this instance it represents a student who obtained a very low score for the Neuroticism factor.

Table 5.2 Descriptive statistics (N=31)

Personality Factor	Minimum	Maximum	Mean	Std. Deviation	Median
Extraversion	0	86	43.48	26.27	43
Agreeableness	12	99	50.39	25.14	41
Conscientiousness	16	87	47.52	19.56	45
Neuroticism	0	94	50.13	23.46	50
Openness to Experience	0	84	32.26	24.20	25

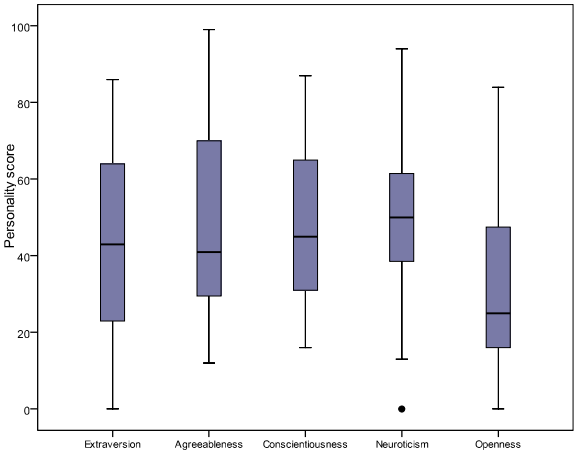


Figure 5.1 Distribution of FFM scores

5.6.1 Correlational Analysis

In this pilot study, students' academic achievement was measured based on their performance in assignments, a midterm test, and final exam. Apart from these individual assessments, pair performance was also measured based on the scores obtained from the tutorial exercises although these scores did not count towards the final course grade. Table 5.3 presents the correlation between students' academic performance and the FFM variables measured using Pearson's correlation coefficients.

Of the five personality factors, only Conscientiousness was found to have a correlation with assignment scores (r^2 = 0.153 and r is 0.39), thus suggesting that Conscientiousness may be a candidate factor influencing academic performance; nevertheless the cause-and-effect of this relationship needs to be empirically tested and validated. The findings relating to the association between Conscientiousness and students' performance were also consistent with studies reported from the psychology literature, i.e. conscientious students tend to perform better, probably due to their positive attributes such as being diligent, organized, and being achievement-oriented (Busato et al., 2000; Duff et al., 2004).

Table 5.3 Correlation between academic performance and the FFM traits (N=31)

	Assign	Test	Final	Extrav.	Agreeab.	Consc.	Neuro.
Assign	1						
Test	.363*	1					
Final	.371*	.355*	1				
Extrav.	-.128	-.334	-.332	1			
Agreeab.	.157	-.003	.074	.357*	1		
Consc.	**.391***	-.054	-.080	.265	.532**	1	
Neuro.	.131	.230	.173	-.369*	-.649**	-.504**	1
Openn.	.142	-.058	.222	.409*	.637**	.208	-.394*

**. Correlation is significant at the 0.01 level (1-tailed).
* Correlation is significant at the 0.05 level (1-tailed).

5.6.2 Pair Performance on Tutorial Exercises

Given that a relatively small sample was available regarding the students' personality profile attending tutorials, and that they were not adequately representing the class population, it was not feasible to perform a statistical testing to assess the effects of personality differences (same or mixed) towards students' academic achievement (i.e. measured by exercise scores). Pair performance based on the scores from the tutorial exercises is shown in Table 5.4. In e*xp1*, pairs of mixed personality showed better performance compared with their counterparts. However, in e*xp2* two pairs from the similar personality group obtained higher scores than another two pairs from the mixed personality group. These comparisons however were not able to be evaluated based on statistical significance; therefore we refrain from making any objective conclusion based on this observation. The exercises given during the tutorial in *exp1* were related to designing an entity relationship (ER) diagram, whereas in *exp2* students were given tasks to convert the ER diagram into a relational model. During *exp3*, students were exposed to developing a JDBC application such as setting up a database

connection and executing SQL queries. Due to the constraints on the duration of the tutorials, students were not able to complete the exercises in *exp3*.

Table 5.4 Tutorial scores per experiment

Experiment		Score (Full point=15)	Pair (Same/Mixed/Unknown)			
			Same personality	Mixed personality	Unknown personality	Total
Exp1	Pair 1	8.0	0	1	-	1
	Pair 2	9.0	1	0	-	1
	Pair 3	13.0	0	1	-	1
	Pair 4	14.0	0	1	-	1
	Pair 5-7	-	-	-	3	3
Total			1 pair	3 pairs	3 pairs	7 pairs
Exp2	Pair 1	5.0	0	1	-	1
	Pair 2	7.0	0	1	-	1
	Pair 3	9.0	1	0	-	1
	Pair 4	9.0	1	0	-	1
	Pair 5	9.0	0	1	-	1
	Pair 6-8	-	-	-	3	3
Total			2 pairs	3 pairs	3 pairs	8 pairs

Table 5.5 shows the matrix correlation between tutorial scores and the FFM for e*xp1*. Analysis indicates that students' performance in their tutorial is positively correlated with the students' Openness to experience (r^2 = 0.329, r =0.574). This result was consistent with the findings reported by Farsides and Woodfield (2003), where academic success was found to be positively associated with Openness to experience. Being Open to experience is believed to provide academic advantage in terms of being critical and analytical in their learning strategies (Farsides & Woodfield, 2003).

Table 5.5 Correlations between tutorial scores and the FFM traits (exp1)

	Tutorial scores	Consc.	Extrav.	Agreeab.	Neuro.	Openn.
Tutorial scores	1					
Conscientiousness	-.019	1				
Extraversion	.241	.378	1			
Agreeableness	.118	.334	.380	1		
Neuroticism	-.256	-.580*	-.414	-.749**	1	
Openness to experience	**.574***	.050	.351	.426	-.324	1

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

The correlation between tutorial scores and personality traits for e*xp2* is shown in Table 5.6. Analysis shows that Extraversion has a negative relationship with tutorial scores (r^2=0.319 and r is -0.565). Both findings in exp1 and exp2 were obtained based on a few tutorial sessions that took place during only two weeks.

Table 5.6 Correlation between tutorial scores and the FFM traits (exp2)

	Tutorial Scores	Consc.	Extrav.	Agreeab.	Neuro.	Openn.
Exercise Scores	1					
Conscientiousness	-.253	1				
Extraversion	**-.565***	.512	1			
Agreeableness	-.119	.564*	.336	1		
Neuroticism	.290	-.710**	-.373	-.705**	1	
Openness	.180	.251	.396	.375	-.144	1

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

5.6.3 Results on Satisfaction and Confidence

Apart from measuring students' academic performance, the pilot experiment also surveyed PP's effectiveness based on the students' level of enjoyment, satisfaction and confidence when pairing. These data were gathered using a questionnaire distributed at the end of each tutorial session. Participants answered questions using a five-point scale ranging from "strongly disagree" to "strongly agree". Figure 5.2 and 5.3 show the distribution of responses for the following questions:

Q1: I felt accomplished working with my partner.

Q2: I enjoyed working with my partner.

Q3: My motivation level increased when working with my partner.

Q4: I understand the topic better when working with my partner.

The responses showed a tendency towards a positive experience during the PP sessions. In both e*xp1* and ex*p2*, on average 24 (82.8%) out of an average of 29 students responded that they agreed (agree and strongly agree) that they felt accomplished working with a partner. On average 27 (93.1%) out of 29 students agreed that they enjoyed the pairing activities. In terms of their motivation to pair, on average 24 (82.8%) out of 29 students responded that their motivation increased when working with a partner. Finally, on average 25 (86.2%) out of an average 29 students agreed that PP helps them understand better the topic.

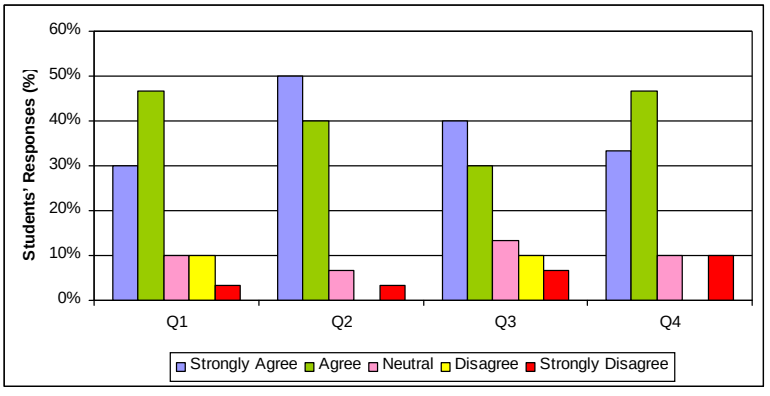


Figure 5.2 PP surveys (exp1)

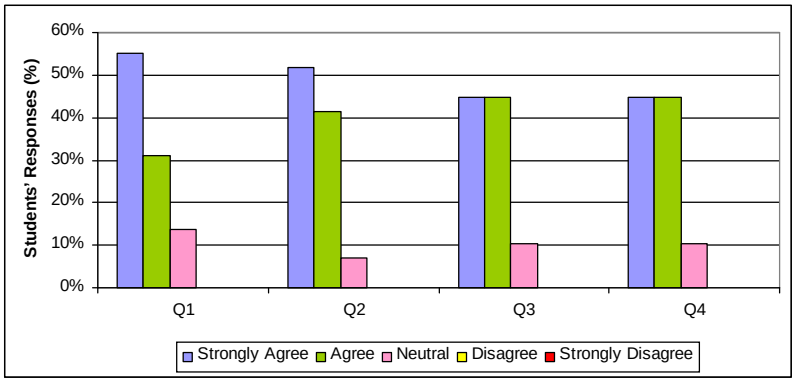


Figure 5.3 PP surveys (exp2)

When asked about the level of satisfaction they had while working in pairs, students on average were highly satisfied for both experiments (see Figure 5.4). Students also responded that the level of confidence in solving the exercise was generally high (see Figure 5.5). On average, approximately 88.5% students indicated that they were satisfied working with their partner and nearly 80% responded that working in pairs increased their confidence level.

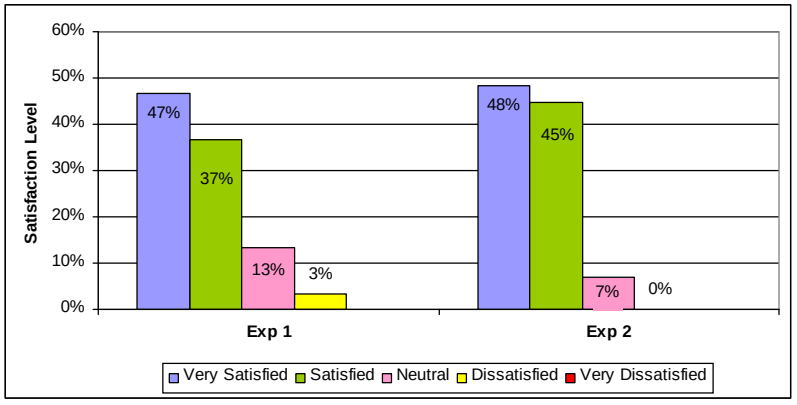


Figure 5.4 Satisfaction level

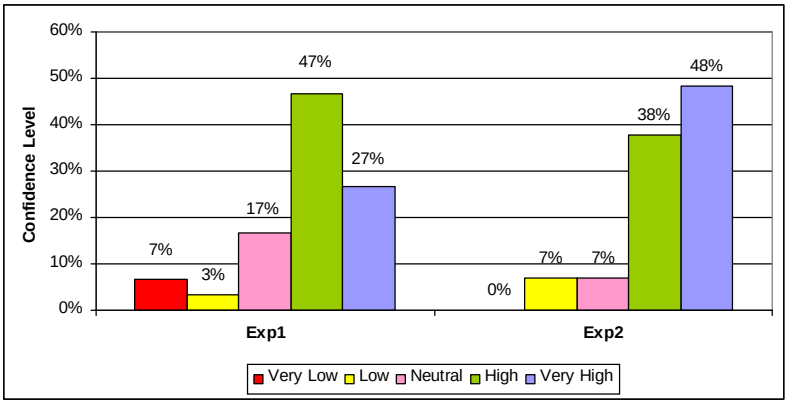


Figure 5.5 Confidence level

5.7 Lessons Learned from the Pilot Experiment

The pilot experiment was the first formal experiment carried out for the purpose to validate the instruments used, and the research methodology chosen for this research. Regarding the

experimental procedure, the initial arrangement was to ensure students swapped their role every 15-20 minutes. However, due to the short duration of each tutorial session (only one hour), pairs only swapped their roles once per tutorial session.

Based on the participants' responses during the tutorial sessions as well as our own observation when students answered the questionnaire, we have made several amendments to the PP questionnaire in order to improve its clarity. The modification was made based on the comments given by the students. One student commented verbally that she felt the first question in the survey (i.e. stated as "*I felt accomplished working with my partner*") was quite vague. This question was then changed to "*I felt that working with this partner was a productive experience".* We also removed the last two questions which required students to rank the factors that would hinder the motivation to pair with another student, so to reduce the complexity of the questionnaire and to allow students to focus their response on the satisfaction and confidence aspects of PP. In addition, the questionnaire was answered at the end of a tutorial session, so the questions should be set as precise as possible to encourage participants to give their feedback quickly and accurately.

Another amendment was to add a new open-ended item to the questionnaire allowing participants to state their comments if any. This helped to identify any areas of improvement for the research based on comments given by the participants. Finally, in terms of the personality test, in addition to the online version, we found that a hardcopy version had to also be used in order to get a larger number of responses.

5.8 Summary

In summary, the results from the pilot study showed that Conscientiousness correlated positively with assignment scores (r = 0.39), Openness to experience to correlated positively with students' performance in the *exp2*'s tutorial exercises (r = 0.57), and Extraversion to be correlated negatively with students' performance in the *exp3*'s tutorial exercises (r = -0.56). In terms of satisfaction and confidence levels, the results showed that students in general were satisfied with the PP experience where on average 88% of students reported that their satisfaction level was higher when working collaboratively with their partner. Similarly, the majority of paired students (80%) responded that PP increased their confidence level in solving the exercises.

We did not perform a hypothesis testing in the pilot study due to the relatively small sample size and also limited data regarding students' personality profile. Furthermore, the primary purpose of the pilot experiment was to identify if there were any major weaknesses in the experimental procedure and the instruments used in this study. As a result of this pilot experiment, we made some modifications to the PP questionnaire as some items needed to be altered or removed and a new item needed to be added.

Chapter 6

THE FIRST EXPERIMENT

This chapter describes a formal experiment conducted at the University of Auckland during the 2009 Summer School. The subjects participating in the experiment were first year undergraduate students enrolled in an introductory programming course (COMPSCI 101). The purpose and details of the experiment are explained in the following sections. Finally, the results obtained are discussed and the limitations of the study are also identified.

6.1 Experimental Objectives

The objective of this experiment was to improve the effectiveness of pair programming (PP) as a pedagogical tool for CS/SE education by investigating the effects that personality differences among paired students may have on PP's effectiveness. A student's personality profile was measured based on the Five Factor personality model. In particular, this experiment focused on the impact of the personality trait Conscientiousness on paired students' academic performance. The main goals of the investigation were to increase the students' satisfaction and the amount of students' learning. These outcomes are reflected in their survey feedback questionnaires (satisfaction) and academic performance as shown by the final grades they achieved for the course (amount of learning).

6.2 Experimental Context

The formal experiment was conducted during the 2009 Summer School involving first year undergraduate students enrolled in an introductory programming course. The teaching of this course consisted of five weeks of lectures and compulsory weekly tutorials. The course instructor taught the fundamentals of programming topics during the two-hour lectures conducted three times per week. The tutorials were held in a computer lab run by a tutor and a few teaching assistants. During the tutorials, students worked in pairs when solving the programming exercises given by the tutor. Each of the tutorial sessions was treated as an independent experiment where the students' feedback regarding the pairing experience was gathered from every tutorial session. During the course, students learnt about object-oriented programming in Java and created a few medium-size applications as part of their assignments. Students willing to participate in the experiments were required to sign the consent form as to fulfill the ethical requirements of the University of Auckland's Human Participant Ethics Committee which approved this experiment prior to commencement.

6.3 Hypothesis

The formulation of the hypothesis investigated in this experiment, as detailed in Chapter 4, relates to understanding the effects of personality traits on the effectiveness of students' learning when using PP. Based on our systematic literature review (SLR), we found that PP advocates propose that diversity or heterogeneity of personalities within a pair is very likely to improve performance (Salleh et al., 2010). For instance, two empirical studies by Sfetsos et al. (2006) and Choi (2004) report that paired students of different personality types achieved better performance compared with paired students of similar personality types.

Many studies reported in the PP literature have applied the Myers-Briggs Type Indicator (MBTI) in assessing students' personality (e.g. Katira et al., 2004; Layman, 2006, Choi et al., 2008). In our research, we adopted the personality framework based on the Five-Factor Model (FFM). The reasons for choosing the FFM are detailed in Chapters 3 and 4. In this experiment, we were interested in investigating whether the evidence holds true (i.e. that different personality is favorable for PP) when using the FFM. In order to investigate the effect of personality differences on PP's effectiveness, we proposed the following null hypothesis:

H_O: Differences in personality trait Conscientiousness do not affect the effectiveness of students who pair programmed.

which is contrasted by the following alternative hypothesis:

H_A: Differences in personality trait Conscientiousness affect the effectiveness of students who pair programmed.

The FFM consists of five broad personality traits known as *Openness to experience*, *Conscientiousness*, *Neuroticism*, *Extraversion*, and *Agreeableness*. Of the five traits, Conscientiousness is the only personality trait that corresponds to achievement orientation (Barrick & Mount, 1998; Costa & McCrae, 1995). This trait is reported to be the most significant for education and learning (De Raad & Schouwenberg, 1996; Busato et al., 2000). Studies in psychology report that Conscientiousness is positively correlated with academic success. Students who are conscientious tend to perform better academically. This is due to the positive characteristics of Conscientiousness such as diligence, hard-work, and ambition.

In this experiment, differences in personality were operationalized by forming pairs consisting of students with different levels of Conscientiousness as measured by the IPIP-NEO. Table 6.1 shows the categorization of pairs according to personality differences using as basis the Conscientiousness factor. The pairing (C_{High}, C_{High}) denotes that a pair consists of students with very similar Conscientiousness levels (higher scores on Conscientiousness). Meanwhile, ($C_{High,}$ C_{Low}) refers to pairs of very different Conscientiousness level (higher and lower scores on Conscientiousness).

We hypothesized that pairs consisting of mixed Conscientiousness levels would achieve better academic performance compared with pairs of students with similar Conscientiousness levels. Our experiment also looked into the association between students' Conscientiousness

levels and their academic achievement, level of satisfaction and confidence with the given tasks.

Table 6.1 Pair configuration

Similar Conscientiousness levels	Mixed Conscientiousness levels
Pair (C_{Low}, C_{Low})	Pair (C_{Low}, C_{Med})
Pair (C_{Med}, C_{Med})	Pair (C_{Med}, C_{High})
Pair (C_{High}, C_{High})	Pair (C_{Low}, C_{High})

6.4 Variables

Evidence from our SLR showed that measuring PP's effectiveness could be achieved using "academic performance", "technical productivity", "program quality", or "satisfaction" (Salleh et al., 2010). Since our study aimed at facilitating CS/SE students through the practice of PP, the metrics to measure PP's effectiveness were "academic performance" and students' "satisfaction". Hence, Conscientiousness level was our independent variable, and PP's effectiveness and satisfaction our dependent variables. PP's effectiveness was measured using assignments, midterm test scores, and final exam scores, whereas satisfaction and confidence levels were measured using a questionnaire where all questions employed a five-point Likert-scale.

6.5 Experimental Procedure

Our hypothesis was investigated using a "single factor between-group design" as the experimental design (Morgan et al., 2004). This design allows each subject to experience only one condition or group, which means, in a particular tutorial, a student was assigned either to a pair of similar Conscientiousness levels or to a pair of mixed Conscientiousness levels (*control group* = similar Conscientiousness levels, *experimental group* = mixed Conscientiousness levels). Therefore, before the first tutorial (i.e. during the first week of semester), students' personality data were gathered using the online IPIP-NEO test. The results of the personality test were used to allocate partners. For this purpose, the personality scores of Conscientiousness were used to assign students between two different groups of similar or mixed personality (e.g. a student with higher score on Conscientiousness was paired with someone with low score on Conscientiousness to form a pair of mixed personality). The Conscientiousness scores (represented by a numerical value from 0 to 99) are divided into low, medium, and high scores based on the scores' range: low Conscientiousness: lowest 40%; medium Conscientiousness: middle 30%; high Conscientiousness: highest 30%.

Every tutorial lasted for two hours. During this time, the tutor explained a topic for about 45 minutes, while the students spent the remaining 75 minutes doing exercises. To allow for "pair-jelling", students worked with their partners for an initial period of 30 minutes; and then

swapped their roles every 15-20 minutes. Before the end of every tutorial, students provided feedback on working with the partner by filling out a questionnaire. The exercises given during the tutorials were graded, thus contributing towards the students' final grade. In addition, assignments and a midterm test were also graded but completed individually. The instruments used in this experiment are detailed in Chapter 4.

The outcomes measured from the experiment were the students' academic performance in their midterm test, three assignments and the final exam scores. Since tutorials varied from week to week, the experiments were designed in such a way so as to minimize the confounding factors which might occur due to differences in tasks and levels of complexity of exercises assigned to the students. Therefore, the tasks and exercises remained the same throughout the week.

6.6 Results and Analysis

This section describes the results of this experiment including the subjects' demographic data. The interpretation of results is presented under the discussion section and finally the potential threats to the validity of the results are also discussed.

6.6.1 Demographics

The subjects involved in the formal experiment were 54 undergraduate CS students. Sixty-five percent (65%) of the subjects were male, 35% female. Subjects' age ranged from 19 to 30 years (median = 20 years old). Subjects came from various ethnic backgrounds; the majority being NZ/Pakeha (13 students, 27%) and Chinese (12 students, 25%). Other ethnic groups included South Korean, Indian, Asian, Middle Eastern, and Pacific Islanders. Of the 32 students who responded to the demographics survey, 84.4% indicated that they did not have any previous work experience. Programming competency was assessed by a survey asking the students to rate their competency on a scale from 1 (very poor) to 5 (outstanding). Twenty-one out of 32 (65.6%) students perceived their programming competency to be poor, 11 (34.4%) students perceived to be at least fair or good. Four students dropped out from the course, thus, they were excluded from our analysis. Of 50 students, only 48 students completed the personality test. Thus, the sample size used in the analysis was 48.

6.6.2 Data Distribution

Figure 6.1 shows the distribution of students' personality scores based on the personality test results. The box in the boxplot represents the middle 50% of the scores, with the upper and lower tails indicating the 75^{th} and 25^{th} percentiles, respectively. The line drawn across the box shows the median value. The boxplot indicates that the median score for Neuroticism is considerably higher than the other factors, whereas the median score for Openness to experience is the lowest. The median score for Conscientiousness is medium (medium levels of Conscientiousness indicate that subjects are reasonably reliable, self-controlled and organized). Except for the Openness to experience trait, in general the spread or variation of

scores between personality factors was quite similar. The greatest spread can be seen in Extraversion.

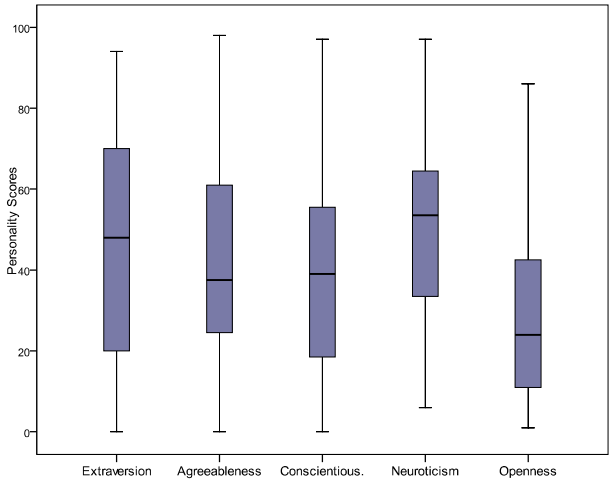


Figure 6.1 Distribution of FFM scores

The distribution of assignment scores between pairs of similar and mixed Conscientiousness levels can be seen in the boxplot shown in Figure 6.2. The boxplot shows that the distribution of assignment scores for the two groups is very similar. Students obtained higher assignment marks regardless of the personality differences in their pairing experience. Notice that there were some outliers in the boxplot. One of the outliers represented a student who did not submit any of the assignments (zero scores), and another who completed the assignments only partially.

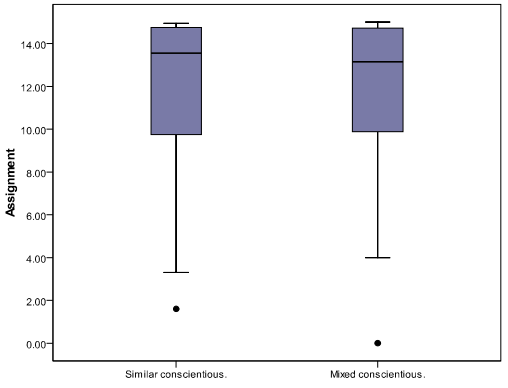


Figure 6.2 Comparison of assignment scores between personality groups

Figure 6.3 shows the distribution of midterm test scores between the two personality groups. The data distribution was negatively skewed for both groups. However, paired students from a mixed Conscientiousness group obtained higher median marks than their counterparts. This indicates that students from the mixed Conscientiousness group performed better in their midterm test compared with students from the similar Conscientiousness levels group.

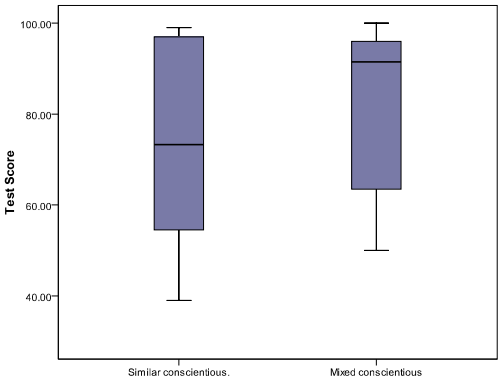


Figure 6.3 Comparison of test scores between personality groups

In terms of individual achievement in the final exam, the dispersion of scores between both groups was similar and the data distribution was also negatively skewed (see Figure 6.4). Similar to the results in the midterm test, the median score for the mixed Conscientiousness group was higher than the similar Conscientiousness group, thus showing that students' performance in the final exam was much better for the former group.

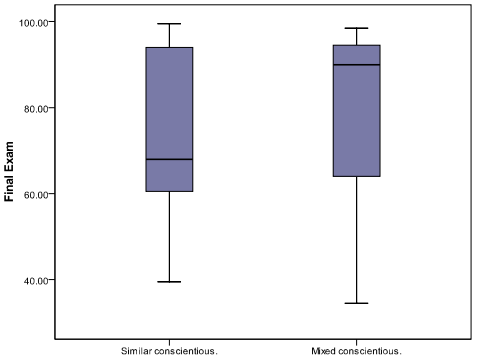


Figure 6.4 Comparison of final exam scores between personality groups

Figure 6.5 shows three boxplots of midterm test scores for each level of Conscientiousness. The distributions of scores between the boxplots have a similar spread, but the median scores for students of low Conscientiousness outperformed the other two groups (medium and high). We noticed that some of the students from the low Conscientiousness group had several years of work experience and reported as having greater perceived programming competency than their peers.

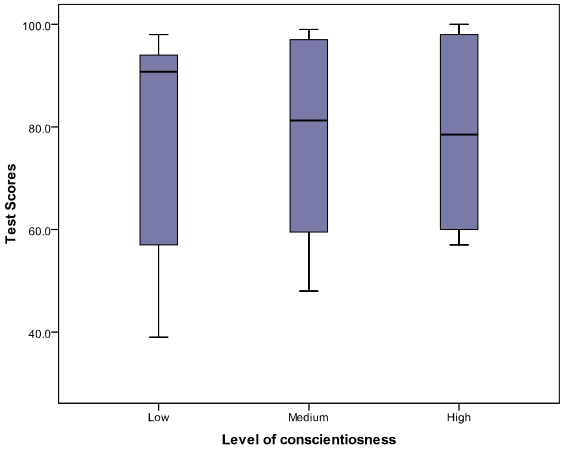


Figure 6.5 Comparison of midterm test scores between Conscientiousness levels

Figure 6.6 shows three boxplots of final exam scores for each level of Conscientiousness. The high Conscientiousness group has the greatest spread or variation of scores. However, the low Conscientiousness group showed higher median scores compared with the other two groups thus indicating that this group also performed better in the final exam than the other groups.

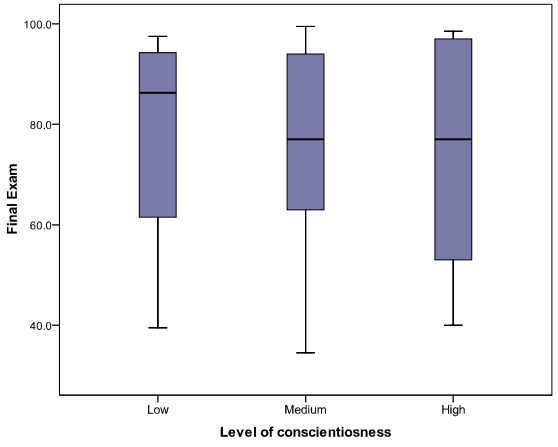


Figure 6.6 Comparison of final exam scores between Conscientiousness levels

6.6.3 Correlational Analysis

In assessing the relationship between variables, one can measure the strength of a relationship using a correlation test (Morgan et al., 2004). Table 6.2 provides the results of the Pearson's correlation between the five personality factors and students' academic performance (assignments, midterm test and final exam scores). Conscientiousness and Openness to experience were the two traits that showed positive correlation with students' performance, but the results were mixed. Conscientiousness showed a positive association with assignments' scores (r=0.29), but no correlation with midterm test and final exam. This result indicates that highly conscientious students typically scored higher marks for their assignments regardless of their pairing configuration. There is also a positive correlation (r=0.34) between final exam scores and Agreeableness. The only personality factor that had

a significant correlation with both the midterm test and final exam scores was the Openness to experience (r^2=0.12, and r is 0.35 for midterm test; r^2=0.08, and r is 0.29 for exam scores). This finding corroborates that of another study (Farsides & Woodfield, 2003) which reported that Openness to experience was positively correlated with undergraduate academic success.

Table 6.2 Correlation between academic performance and personality factors (N=48)

	Assign	Test	Final	Extrav.	Agreeab.	Consc.	Neuro.
Assign	1						
Test	0.36**	1					
Final	0.42**	0.88**	1				
Extrav.	-0.05	0.08	0.04	1			
Agreeab.	-0.01	0.19	**0.34***	0.10	1		
Consc.	**0.29***	0.07	-0.05	0.27*	0.14	1	
Neuro.	-0.17	-0.04	-0.03	-0.49**	-0.07	-0.53*	1
Openn.	0.54	**0.35***	**0.29***	0.35**	0.20	-0.08	0.06

**. Correlation is significant at the 0.01 level (1-tailed).
*. Correlation is significant at the 0.05 level (1-tailed).

6.6.4 Hypothesis Testing

We used a single factor multivariate analysis of variance (MANOVA) to analyze whether there was any significant difference in academic achievement between paired students of similar and mixed Conscientiousness level. MANOVA is regarded as a complex statistic that linearly combines several dependent variables in a single analysis, where variables need to be correlated at a low to moderate level (Leech et al., 2005). Herein, assignments, test and final exam scores were analyzed simultaneously using the General Linear Model program in SPSS.

Table 6.3 provides mean values and standard deviation values for assignments, test and final exam scores, for each group. Mean differences are almost the same for assignments' scores but somewhat different for the midterm test and final exam scores. The Levene's Test (see Table 6.4) indicates that the assumption of homogeneity of variances of each variable was not violated.

Table 6.3 Mean and standard deviation of paired students of similar and mixed Conscientiousness

	Personality Type	**Mean**	**SD**	**N**
Assignments	Similar Conscientiousness	13.07	2.08	22
	Mixed Conscientiousness	12.48	2.53	21
	Total	12.78	2.30	43
Test Scores	Similar Conscientiousness	76.00	20.68	22
	Mixed Conscientiousness	83.83	16.21	21
	Total	79.83	18.83	43
Final Exam	Similar Conscientiousness	73.11	18.68	22
	Mixed Conscientiousness	78.21	22.00	21
	Total	75.60	20.29	43

Table 6.5 shows the results for differences on performance between the two groups. MANOVA generated four multivariate tests (by default). Of these four tests, the one that provides "*good and commonly used multivariate F*" is Wilks' Lambda (Leech et al., 2005). Thus, referring to Wilks' Lambda (under the "PairType" effect), results showed no significant

differences (*F*=1.03, *df*=39, *p*=0.39) between the "PairType" groups, on a linear combination of three dependent variables (assignments, test and final exam scores). Thus, using the 95% confidence interval we failed to reject the null hypothesis based on our data, thus supporting the view that heterogeneity of personality traits did not affect the effectiveness of students who pair programmed.

Table 6.4 Levene's tests

DV	F	df1	df2	Sig.
Assignments	1.13	1	41	0.29
Test Scores	3.58	1	41	0.07
Final Exam	0.40	1	41	0.53

Table 6.5 Multivariate tests

Effect	Test approach	Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	0.98	746	3.0	39.0	0.0
	Wilk's Lambda	0.02	746	3.0	39.0	0.0
	Hotelling's Trace	57.44	746	3.0	39.0	0.0
	Roy's Largest Root	57.44	746	3.0	39.0	0.0
PairType	Pillai's Trace	0.07	1.03	3.0	39.0	0.39
	Wilk's Lambda	**0.93**	**1.03**	**3.0**	**39.0**	**0.39**
	Hotelling's Trace	0.79	1.03	3.0	39.0	0.39
	Roy's Largest Root	0.79	1.03	3.0	39.0	0.39

6.6.5 Statistical Power Analysis

Although the purpose of conducting hypothesis testing is to identify whether or not one should reject the null hypothesis (i.e. the ability to detect that effects do exist), there is a possibility of making an incorrect decision (Gravetter & Wallnau, 2004). This is due to our limitations in making a general conclusion about the entire population when we are only able to infer a pattern based on the sample size available (Gravetter & Wallnau, 2004). Gravetter & Wallnau (2004) mentioned that "*there is always a chance that the sample is misleading and will cause a researcher to make the wrong decision about the research results*" (p. 188). The two types of errors in a hypothesis test are known as *Type 1* and *Type 2* error.

The Type 1 error (termed α) occurs when the null hypothesis is rejected when in fact it is true. This means that there is a statistically significant difference found between the treatment groups when in fact no true difference exists in the entire population. The Type 2 error (termed β) occurs when the hypothesis testing does not reject the null hypothesis although a true difference exists between the groups in the entire population (Gravetter & Wallnau, 2004).

In our experiment, we found a lack of support for our alternative hypothesis, thus accepting the null hypothesis. Therefore, there is a risk of committing a Type 2 error. It is vital to ensure that the hypothesis test is making the correct decision and this can be achieved by means of statistical power analysis (Miller, Daly, Wood, Roper, and Brooks, 1997). The power of a statistical test is defined as "*the probability that the H_O will be rejected when it is false*" (Cohen, 1992, p. 98). If the statistical power is high, there is a high probability of obtaining a statistically significant result (i.e. if the effect is truly exists, it will be highly likely detected), whereas low power indicates that a study is inconclusive if the findings are not significant

(Dyba, Kampenes, & Sjoberg, 2006). Thus the power analysis gives us an indication of how much confidence we have in our failure to reject the null hypothesis.

According to Cohen (1988), the power of a statistical test is the probability of rejecting the null hypothesis when it is false (i.e. H_O is correctly rejected). Hence, power is 1 – β. In order to compute the power of a statistical test, three factors needed are: a) S*ignificance level (α)*; b) S*ample size*; and c) *Population effect size* (Cohen, 1988).

The type of power analysis we carried out in this research is known as a *post-hoc statistical power* (Lan & Lian, 2010). This type of analysis was performed because the statistical analysis of our data has been carried out and the power analysis helps explain our findings. We carried out the power analysis using a stand-alone statistical power software package known as *G*Power* (version 3.1.2).[4] G*Power supports power analysis for statistical tests commonly used in social and behavioral sciences research. It has also been applied in many other disciplines such as biology, ecology, pharmacology, and medical research (Faul, Erdfelder, Lang, & Buchner, 2007; Faul, Erdfelder, Buchner, & Lang, 2009). This software was evaluated positively in some reviews (Faul et al., 2007; Dattalo, 2009).

The power analysis was executed based on our multivariate test results. Table 6.6 lists the input parameters involved in the calculation and the resulting power value. These results showed that our obtained power is considered to be low (28%) at the small effect size of 0.08. According to Dyba, Kampenes and Sjoberg (2006), statistical tests in SE experiments should achieve a power level of at least 80% in order to produce a reliable conclusion regarding the acceptance or rejection of the null hypothesis. This means that although the data used did not support the alternative hypothesis, a larger sample size might have shown support for it.

Table 6.6 Protocol of power analyses

F tests – MANOVA: Special effects and interactions		
Options:	Wilks U, O'Brien-Shieh Algorithm	
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size $f^2(U)$	= 0.08
	α err prob	= 0.05
	Total sample size	= 43
	Number of groups	= 2
	Number of predictors	= 1
	Response variables	= 3
Output:	Noncentrality parameter λ	= 3.38
	Critical F	= 2.84
	Numerator df	= 3.0
	Denominator df	= 39.0
	Power (1-β err prob)	= 0.28
	Wilks U	= 0.927

The relationship between statistical power and the sample size is plotted in a graph shown in Figure 6.7. In this graph, the power (1 – β) (y-axis) is the function of the sample size (x-axis). The value of statistical power varies as a function of the effect size, the significance

[4] G*Power 3 is freely available for download from the *Institut fur Experimentelle Psychologie* at *Heinrich Heine-Universitat Dusseldorf* (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3).

level (α) and the total sample size (Cohen, 1988). At the small effect size of 0.08, the power value increased as the total sample size increased. Figure 6.8 shows the power plot at three different α values (from 0.05 to 0.15). In order to obtain a high statistical power of 80%, the sample size required is approximately 140 (for α =0.05), and 100 (for α = 0.15).

As mentioned by Stevens (2002), power is not an issue when the sample size is large (100+ per group). In addition, in order to obtain a greater statistical power, there should be a compromise in setting the significance (α) level (Stevens, 2002). Dyba et al. (2006) also recommend that SE researchers consider setting the α value at a more lenient level rather than using the conventional value of 0.05 in order to balance the probabilities of committing the Type 1 and Type 2 error.

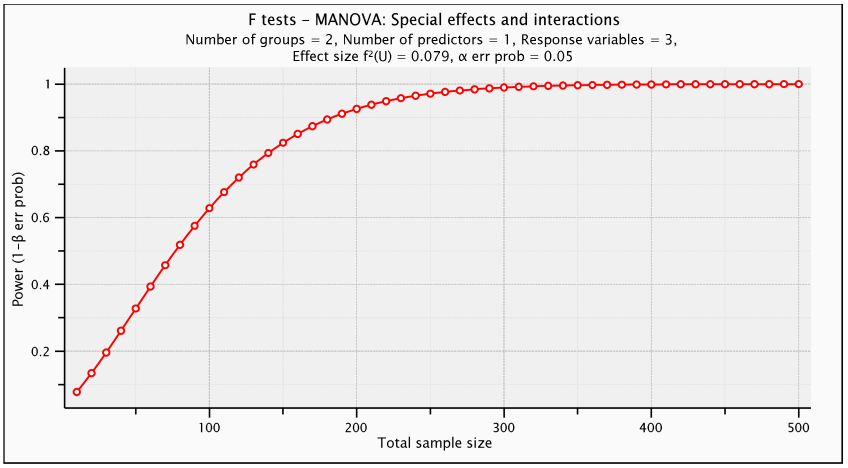


Figure 6.7 Power plot

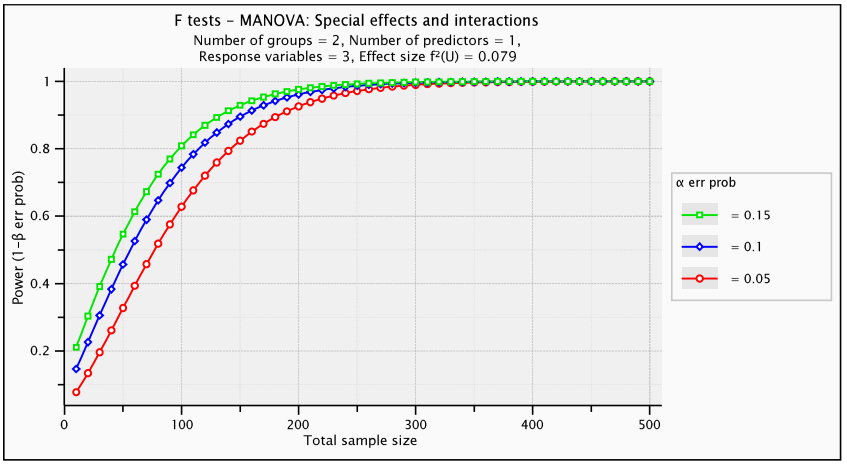


Figure 6.8 Multiple power plot

6.6.6 Results for Satisfaction and Confidence

The response rate of the post-experimental survey was approximately 67% in every tutorial. The surveys were distributed in the second week of the semester until the final week of tutorials (altogether nine tutorials). Data was analyzed separately as each tutorial was treated

as a single independent "mini-experiment". Our analysis showed that overall students obtained a large degree of satisfaction and confidence from the pairing activity (see Figures 6.9 and 6.10). In terms of satisfaction, on average 88.54% students were satisfied working with their partner, and approximately 87.88% responded that their level of confidence solving the exercises with their partner was high.

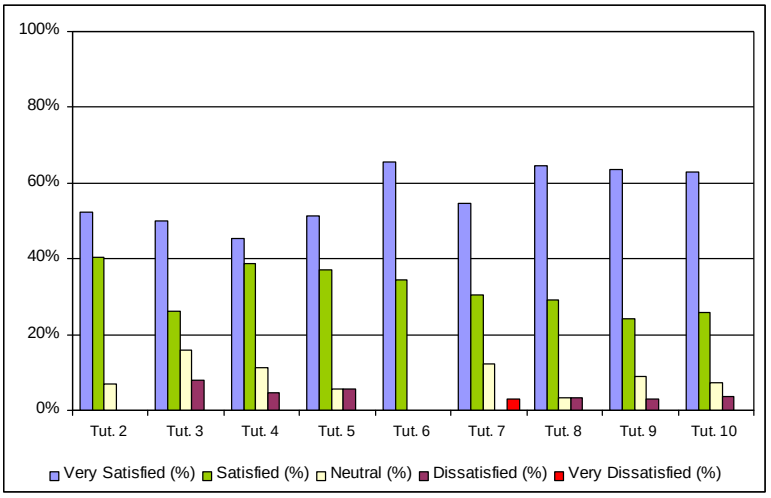


Figure 6.9 Survey on PP satisfaction

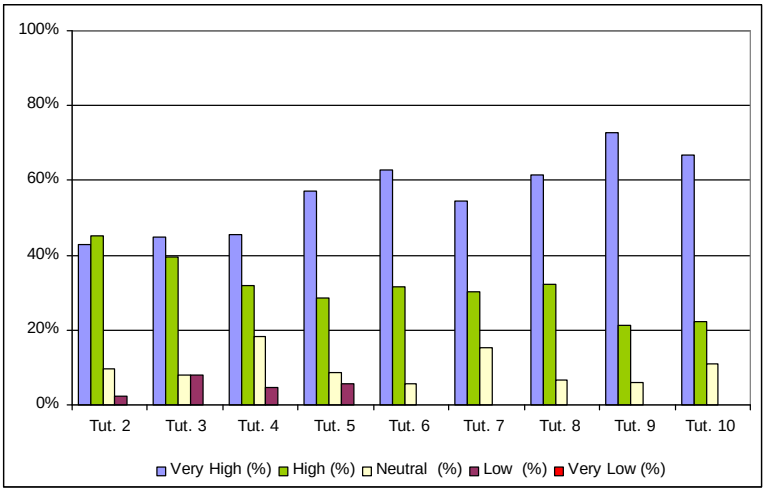


Figure 6.10 Survey on PP confidence

Further analyses on subjects' responses regarding their PP's experience revealed that on average 90.4% students agreed (both agree and strongly agree) that working with their partner was a productive experience. On average, 92.6% students enjoyed working collaboratively with their partner. In terms of students' level of motivation, on average 86% students responded that their motivation increased when working with a partner. Subjects gave responses by answering questions using a five-point Likert scale ranging from "strongly disagree" to "strongly agree". Figure 6.11 – 6.13 show the distribution of responses for the following questions:

Q1: I felt that working with this partner was a productive experience.

Q2: I enjoyed working with my partner.

Q3: My motivation level increased when working with my partner.

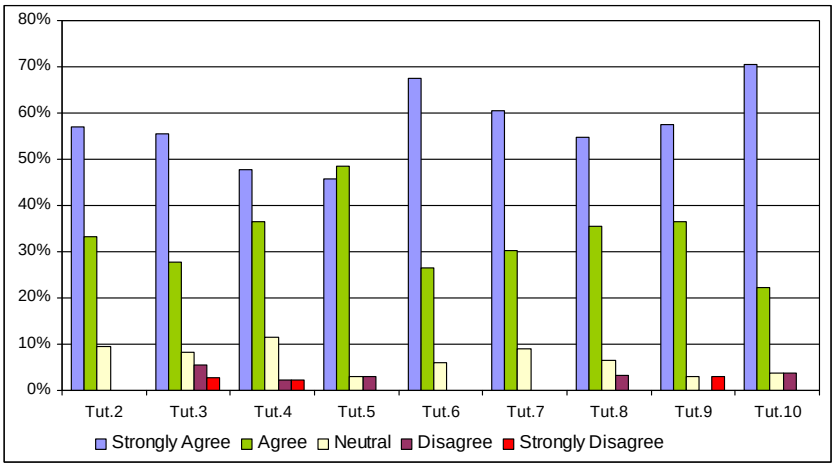


Figure 6.11 Responses on PP's experience (Q1)

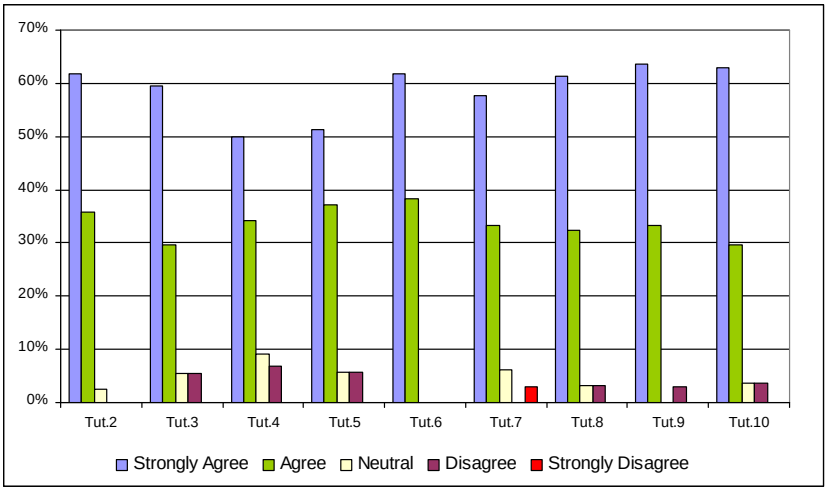


Figure 6.12 Responses on PP's experience (Q2)

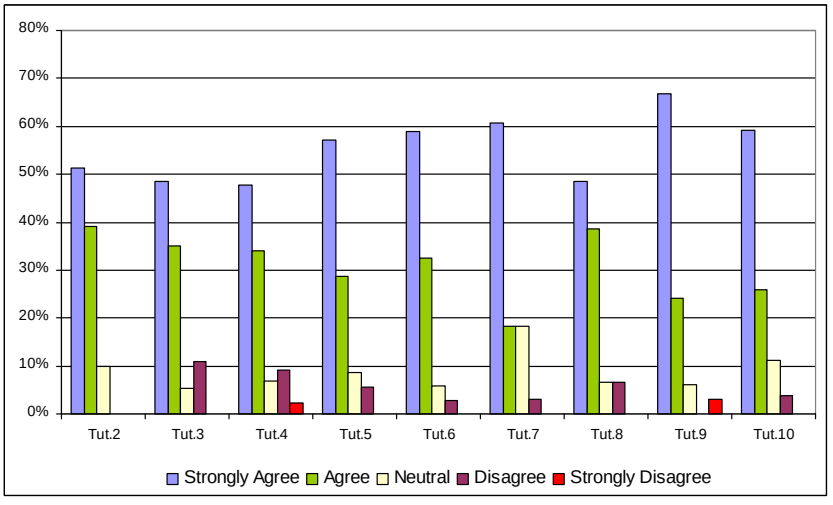


Figure 6.13 Responses on PP's experience (Q3)

To answer the question on whether there are any differences in the level of satisfaction between the controlled and experimental groups, we applied the Mann-Whitney U test to each of the experiments' unit. Nonparametric testing was chosen because the dependent variable (i.e. satisfaction level) was not normally distributed thus violating the assumptions for parametric testing. In Table 6.7, the mean satisfaction ranks for paired students are shown. The group with the highest mean rank had the highest level of satisfaction.

Although the similar Conscientiousness group appeared to score higher ranks in most of the experiments, these differences were not always significant. As can be seen in Table 6.8, using a significance level of 0.05, there were no significant differences between groups, except for the Tut. 10. Overall, results demonstrated that the satisfaction levels of paired students were not affected by personality differences based on Conscientiousness, and paired students achieved higher satisfaction regardless of their differences in personality when pairing.

Table 6.7 Mann-Whitney U Ranks for satisfaction level

Tutorials	Pair Type	N	Mean Rank	Sum of Ranks
Tut. 2 (N=39)	Similar Conscientiousness	13	23.46	305.0
	Mixed Conscientiousness	26	18.27	475.0
Tut. 3 (N=37)	Similar Conscientiousness	23	21.09	485.0
	Mixed Conscientiousness	14	15.57	218.0
Tut. 4 (N=36)	Similar Conscientiousness	22	17.55	386.0
	Mixed Conscientiousness	14	20.00	280.0
Tut. 5 (N=26)	Similar Conscientiousness	15	14.43	216.5
	Mixed Conscientiousness	11	12.23	134.5
Tut. 6 (N=34)	Similar Conscientiousness	9	18.17	163.5
	Mixed Conscientiousness	25	17.26	431.5
Tut. 7 (N=31)	Similar Conscientiousness	10	16.30	163.0
	Mixed Conscientiousness	21	15.86	333.0
Tut. 8 (N=30)	Similar Conscientiousness	13	18.31	238.0
	Mixed Conscientiousness	17	13.35	227.0
Tut. 9 (N=31)	Similar Conscientiousness	15	17.60	264.0
	Mixed Conscientiousness	16	14.50	232.0
Tut. 10 (N=24)	Similar Conscientiousness	13	15.50	201.5
	Mixed Conscientiousness	11	8.95	98.5

Table 6.8 Mann-Whitney U test statistics for satisfaction level

Tutorials	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)	Exact Sig.
Tut. 2	124.0	475.0	-1.51	0.13	0.19
Tut. 3	113.0	218.0	-1.62	0.11	0.14
Tut. 4	133.0	386.0	-0.75	0.45	0.51
Tut. 5	68.5	134.5	-0.79	0.43	0.47
Tut. 6	106.5	431.5	-0.28	0.78	0.82
Tut. 7	102.0	333.0	-0.14	0.89	0.92
Tut. 8	74.0	227.0	-1.80	0.07	0.13
Tut. 9	96.0	232.0	-1.07	0.29	0.36
Tut. 10	32.5	98.50	-2.63	0.01	0.02

6.7 Discussion

We observed an interesting finding when investigating the relationship between paired students' performance and personality traits. We found that the personality traits that appeared to have a positive correlation with academic performance were Conscientiousness, Agreeableness, and Openness to experience. This result is in line with several existing studies in psychology (Chamorro-premuzic & Furnham, 2008; Farsides & Woodfield, 2003; Conard, 2006), and business (Nguyen et al., 2005). They found that high Conscientiousness students were more likely to perform well in the class compared with low Conscientiousness students. Likewise, the Agreeableness and Openness to experience traits are also reported to have significant positive correlation with academic success (Farsides & Woodfield, 2003; Chamorro-premuzic & Furnham, 2008).

Our failure to support the alternative hypothesis can be attributed to several reasons. Based on the statistical power analysis, we found that there is a low power in our statistical testing. The power ($1 - \beta$) indicates that the possibility of detecting a difference between the groups was only of 28%. In future studies, power perhaps could be increased by increasing the sample size. Stevens (2002) asserts that a test's power is heavily dependent on the sample size; an increase in a sample size causes the power to increase dramatically. Dyba et al. (2006) also recommend that increasing the sample size is one of the ways to increase the statistical power in SE experiments. This is because the ability to detect differences across the studied groups is greater in a larger sample size (Dyba et al., 2006; Cook & Campbell, 1979).

There is also an issue with regard to the process of forming a pair of similar or mixed personality. For example, matching a high conscientious with a low conscientious student can possibly produce an incompatible pair due to dissimilarities in character and attitude (based on the survey, we noted comments from students in mixed personality pairs who did not enjoy working with their partner). Likewise, forming a pair of similar personalities where both are low conscientious students may bring disadvantages to the pair due to their lack of self-discipline and low need for achievement. In this sense, comparing the performance of paired students between similar and mixed personalities had a few issues.

There is also a possibility that the performance may be affected by gender differences, as reported by Nguyen et al. (2005) in their investigation about the moderating role of gender in determining the relationship between personality and academic performance. In their study, they found that emotional stability (the reverse of the Neuroticism factor) positively and significantly predicted academic performance of male students, but the same prediction did not occur for female students.

A large and growing body of literature has investigated the effect of personality composition towards team effectiveness (Mohammed & Angell, 2003; Peeters et al., 2006). In one of the meta-analytic studies, Bowers et al. (2000) investigated whether the teams that were homogeneous in personality outperformed the teams consisting of heterogeneous personalities and the findings showed a partial support for heterogeneous teams. They also

suggested that effective team personality composition was highly dependent on the type of task, the difficulty level, and the level of communication required in performing the task (Bowers et al., 2000). Other authors also pointed out that task type can play a significant role in determining effective personality composition (Mohammed & Angell, 2003; Driskell et al., 2006). While these studies were conducted mostly in the psychology and business domains, further research should be done to investigate the personality composition affecting PP's effectiveness as a pedagogical tool. The issue of whether homogeneity or heterogeneity of personality is good for PP has not been clearly solved yet.

Another possible explanation for our results to support the null hypothesis might be related to other confounding factors such as the skill level of paired students. It would be useful to investigate which factor was the strongest predictor of PP's effectiveness; for instance, by investigating the correlation between personality traits and skill levels towards performance of paired teams. This is because personality might not be a strong predictor of PP success when compared with the skill level among paired students. Results from our systematic review revealed that PP worked best when the skill level gap between partners was not too broad (Salleh et al., 2010).

Our results also showed that PP helped students achieve high satisfaction and great confidence in learning programming. Overall, 92% of the paired students indicated that they were happy working with their partner. These results are consistent with those of existing studies that investigated PP's effectiveness (Mendes et al., 2005; Mendes et al., 2006; Sfetsos et al., 2006; Williams & Kessler, 2000).

6.8 Threats to the Validity

There are some uncontrolled variables which may have affected the validity of the experimental results. One of these was students' previous programming experience. We noticed that some of the students who already had a few years of programming experience achieved high scores in their test, but scored somewhat low on Conscientiousness. They were strong programmers with appropriate knowledge and know-how of programming compared to other students. Being highly conscientious may not be necessary for these students in order to obtain good academic results in this particular course.

Another threat was with regard to the change of partners during the tutorials due to a partner's absenteeism. Some students failed to turn up to their allocated tutorial and attended a different tutorial without informing the tutor. This created an unbalanced number of pairs between groups and the likelihood that some students in the controlled group to be moved to the experimental group. The small sample size used in this study (48 students) may also have affected the significance of the results. This is in particular due to the low level of power generated based on our statistical testing. Cook & Campbell (1979) mentioned that when the sample size is large enough, even very small effects can be statistically significant.

6.9 Summary

The focus of this study was to determine whether differences in students' personality profiles during pairing activities would impact their academic performance. The results of the formal experiment showed a positive significant correlation between some specific performance measures and the three personality factors: i) Conscientiousness was positively correlated with assignments' scores (r=0.29); ii) midterm test and final exam marks were positively correlated with Openness to experience (r=0.35, and r=*0.29* respectively); and iii) Agreeableness was positively correlated with final exam marks (r=0.34).

The current study did not reject the null hypothesis, thus it did not provide any evidence for distinguishing the performance of paired students between similar and mixed Conscientiousness levels (p = 0.93, CI=95%). The low level of power generated from the statistical test indicates that our experiment had a poor power for detecting the difference, which may result from the small effect size and/or small sample size used in this study.

On average, 88% of students were satisfied with the PP experience. Similarly, most of the students (87%) responded that their confidence level increased when working in pairs. The evidence from this study suggests that regardless of the variation in students' personality disposition, PP not only caused the increase of satisfaction and confidence level, but also brought enjoyment to the class and enhanced students' motivation for learning.

In summary, the current findings add to our understanding of the effect of personality variation towards students' academic performance when practicing PP. One of the major implications is to further investigate personality traits of paired students focusing on other personality factors such as Openness to experience, Agreeableness, and Neuroticism. These factors are reported as educationally relevant (De Raad & Schouwenburg, 1996). The next chapter describes the experiment conducted in the subsequent semester using the same course and instrumentation.

Chapter 7

THE SECOND EXPERIMENT

This chapter describes the second experiment of this research programme, conducted at the University of Auckland during the First Semester of 2009. The subjects who participated in the experiment were first year undergraduate students enrolled in an introductory programming course (COMPSCI 101). The purpose and details of the experiment are explained in the following Sections. Finally, the results obtained are discussed and the limitations of the experiment are also identified.

7.1 Experimental Objectives

The formal experiment described in this chapter is an extension of the experiment described in Chapter 6, where PP's team personality composition was investigated. The previous experiment investigated how differences in paired students' personality profiles affected their academic performance, focusing on the FFM's Conscientiousness factor. Conscientiousness was chosen, and also used in the present experiment, because this factor was reported to be associated with team performance as well as academic success (Chamorro-premuzic & Furnham, 2003b; Fruyt & Mervielde, 1996; Duff et al., 2004).

The results from our previous experiment (in 2009 Summer School) did not provide strong support to distinguish performance between paired students of similar and mixed personalities. However, despite its small sample size, one of the important issues that came to light as a result of that work related to the pair formation strategy: paired students of mixed personality consisted of students of high and low Conscientiousness and such a matching could possibly produce an incompatible pair due to dissimilarities in character or attitude (Kichuk & Wiesner, 1997). Thus, in the present experiment we used a different approach to pair formation in order to overcome such issues.

This formal experiment's objective is to improve the effectiveness of PP as a pedagogical tool for CS/SE education by investigating the influence of the FFM personality model's Conscientiousness factor towards pairs' academic performance. Each subject's personality profile was measured based on the Five Factor personality model. The main objectives of the investigation are to increase students' satisfaction and the amount they learn. These outcomes are reflected in students' academic performance as shown by their performance in assignments, a midterm test and a final exam. These research objectives were outlined using the Goal/Question/Metric (GQM) framework (Basili et al., 1999) shown in Table 7.1. The detailed goal definition for the experiment is as follows:

Object of study: PP technique.

Purpose: To improve the effectiveness of PP as a pedagogical tool in higher education institutions.

Focus: To investigate the influence that the FFM personality model's Conscientiousness factor can potentially have over the success of the PP practice in CS/SE courses/tasks.

Perspective: From the point of view of the researcher.

Context: In the context of undergraduate CS/SE students.

Table 7.1 GQM definition

Goal(s)	Question(s)	Metric(s)
To investigate the effect of Conscientiousness towards successful pair configuration.	Do differences in Conscientiousness level within a pair affect the pair's academic performance?	Students' academic achievement measured by assignments, midterm test and final exam scores.
To investigate the level of satisfaction of paired students.	Did students feel satisfied working in pairs?	PP questionnaire on satisfaction level.
To investigate the level of confidence of paired students.	Did students feel confident working in pairs?	PP questionnaire on confidence level.

7.2 Experimental Context

The formal experiment was conducted during the first semester of 2009, where participants were first year undergraduate students enrolled in an introductory programming course (COMPSCI 101). In this course the fundamentals of programming and an introduction to the programming language Java were taught by an instructor. The teaching component of this course consisted of twelve weeks of lectures (36 lectures, each lasting for one hour) and ten weeks of compulsory tutorials (each lasting for two hours). The tutorials were held in a computer lab run by a tutor and a few teaching assistants. During the tutorials, students worked in pairs solving the programming exercises given. Each of the tutorial sessions was treated as an independent experiment where data about the students' pairing experience was gathered from every session. Students were given three assignments throughout the course, which involved developing small to medium-size applications. Students willing to participate in the experiments were required to sign a consent form to fulfill the ethical requirements of the University of Auckland's Human Participant Ethics Committee.

7.3 Hypothesis

According to the FFM, the level of Conscientiousness a person has indicates the degree of aspiration or one's desire for achievement (McCrae & John, 1992). Therefore, a highly conscientious individual tends to be more persistent, responsible, organized, thorough, and ambitious. In contrast, persons with low Conscientiousness are expected to be impulsive, irresponsible, disordered, and to lack a desire for achievement (Driskell, Hogan, & Salas, 1987).

Research on personality suggests that students' academic performance is positively associated with their level of Conscientiousness (Busato et al., 2000; Zyphur, Bradley, Landis, & Thoresen, 2008; Chamorro-premuzic & Furnham, 2003b). It has also been reported in the literature that teams comprising a higher average score of Conscientiousness demonstrated better job performance (Barrick et al., 1998; Neuman et al., 1999). Extending this logic into the realm of PP, we predicted that pairs consisting of highly conscientious students are expected to achieve better academic performance than pairs consisting of students with low levels of Conscientiousness. Hence, the following null hypothesis was investigated in our experiment:

H_O: Differences in Conscientiousness level do not affect the academic performance of students who pair programmed.

which is contrasted by the following alternative hypothesis:

H_A: Differences in Conscientiousness level affect the academic performance of students who pair programmed.

Table 7.2 shows the categorization of pairs according to students' Conscientiousness level. Pair (C_{High}, C_{High}) denotes a combination in which both students have high levels of Conscientiousness. This experiment compared the performance of students in these groups based on their academic achievement in the course. Our experiment also looked into the association between each student's personality score with their academic performance, level of satisfaction and confidence when working in pairs.

Table 7.2 Pair configuration

Conscientiousness Level	Pairing Groups
High	Pair (C_{High}, C_{High})
Medium	Pair (C_{Med}, C_{Med})
Low	Pair (C_{Low}, C_{Low})

7.4 Variables

Our synthesis of evidence from the SLR showed that measuring PP's effectiveness could be achieved using "academic performance", "technical productivity", "program quality", or "satisfaction" (Salleh et al., 2010). Since our experiment aims at facilitating CS/SE students through the practice of PP, the metrics selected to measure PP's effectiveness were "academic performance" and students' "satisfaction". Hence, PP's effectiveness and satisfaction were our dependent variables and level of Conscientiousness our independent variable (single-factor).

In this research, PP's effectiveness was measured using assignments, a midterm test and final exam scores. Level of satisfaction and confidence were measured using a questionnaire where all questions employed a five-point Likert-scale. We used the same instruments as in our previous experiment.

7.5 Experimental Procedure

We followed the same procedure carried out in our previous experiment (see Chapter 6), where each of the tutorial sessions was treated as an independent formal experiment. During the first week of the semester, students' personality data were gathered using the online IPIP-NEO test. The results of the personality test were then used to allocate partners. For this purpose, the numerical personality scores relating to the Conscientiousness factor were used to assign students to one of three possible groups: low, medium or high level of Conscientiousness. Based on the distribution of scores for the Conscientiousness trait, the grouping was done to provide a more balanced number of subjects within each group. As such, the classification of Conscientiousness level used herein is based on the range of scores: lowest 45% (low Conscientiousness); middle 25% (medium Conscientiousness); and the highest 30% (high Conscientiousness).

The allocation of pairs within each group was performed randomly. Since we only had Conscientiousness as our independent variable, our hypothesis was investigated using a "*single factor between-group design*" as the experimental design (Morgan et al., 2004).

Every tutorial lasted for two hours. During this time, the tutor explained a topic for about 45 minutes, and the students completed exercises for the remaining 75 minutes. To allow for "pair-jelling", students worked with their partners for an initial period of 30 minutes; and then swapped their roles every 15-20 minutes. Before the end of every tutorial, students provided feedback on working with the partner by filling out a questionnaire. The exercises given during the tutorials were graded, thus contributing towards the students' final grade. In addition, assignments and a midterm test were also graded but completed individually.

The outcomes measured from the experiment were the students' academic performance in their midterm test, final exam, and three assignments. Since tutorials varied from week to week, the experiments were designed in such a way to minimize the confounding factor which might occur due to differences in tasks and level of complexity of exercises assigned to the students. Therefore, the tasks and exercises remained the same throughout the week.

7.6 Results and Analysis

This section describes the results of the experiment including subjects' demographic data. The interpretation of results is presented under the discussion section and finally the potential threats to validity of the results are also discussed.

7.6.1 Demographics

A total of 453 students enrolled in the COMPSCI 101 course for the first semester of 2009. Of these, 295 (65.1%) planned to obtain a BSc, 44 (9.7%) a BArts, 22 (4.9%) a BCom, 6 (1.5%) a Graduate Diploma of Science degree, and the remaining to obtain a Bachelor of Law co-joint with Commerce and Science.

There were 350 male students (77%), 103 females (23%), and subjects' age ranged from 19 to 52 years (median age = 19 years). Of the 317 students who responded to the demographics survey, 255 students (85%) indicated that they did not have any work experience; however, 30 students (9.5%) indicated their programming competency as above average. Subjects came from various ethnic backgrounds: 102 (22.5%) NZ/Pakeha, 85 (18.8%) Chinese, 25 (5.5%) Indian, 25 (5.5%) Korean and the rest are Pacific Islanders, European, Middle Eastern, African, Asian, and American. Of the 453 students, 212 students (47%) completed the personality test and have taken either the midterm test or final exam. Therefore, the sample size used in our analysis comprised of 212 data points.

7.6.2 Correlation Analysis

The distribution of personality scores based on the FFM test is shown in Figure 7.1. The rectangle in the boxplot represents the middle 50% of the cases, with the upper and the lower lines representing the remaining upper and lower 25%, respectively. As can be seen in Figure 7.1, the distributions of scores between the Agreeableness and Conscientiousness have a similar spread, and their median value is also very similar. The median and the distribution of personality scores between Extraversion and Neuroticism are also very similar. Of the five factors, Openness to experience had the distribution of data most positively skewed. The black dot outside the distribution range is considered an outlier, and in this instance it represents a student who obtained a very high score in the Openness to experience factor.

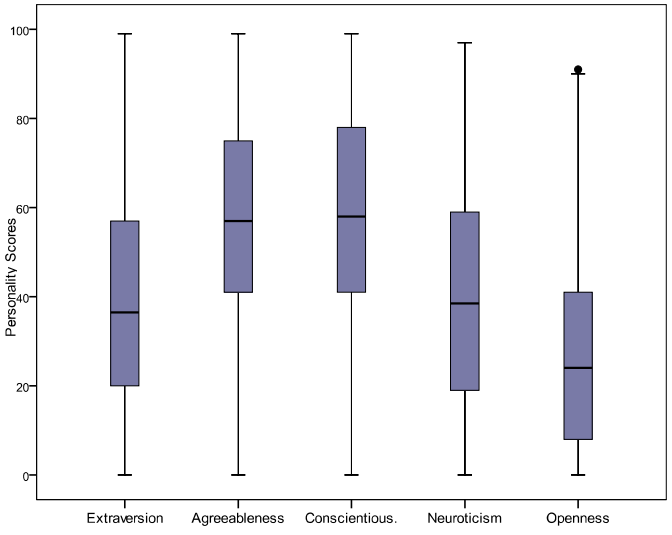


Figure 7.1 Comparison of FFM scores

Figure 7.2 shows the distribution of assignment scores according to students' levels of Conscientiousness. The distribution of assignment scores across the three groups shows a similar spread and they are all negatively skewed. The median value for these groups is also very similar. This means most students obtained high marks on their assignments regardless

of their Conscientiousness level. Note that each group had a few outliers, representing students who did neither complete their assignments nor submit some of the assignments.

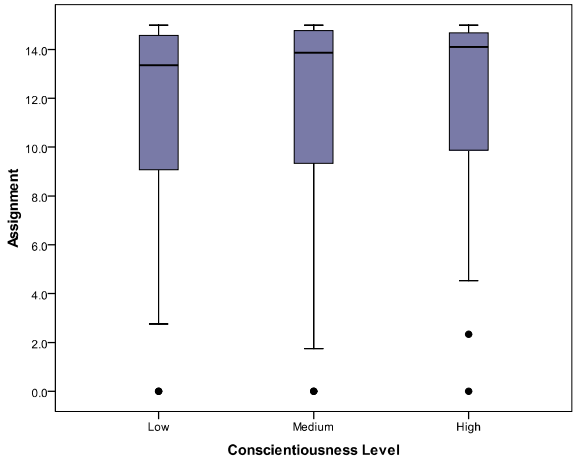


Figure 7.2 Comparison of assignment scores between groups

In terms of students' test scores, the distribution of data for the three groups showed a negative skew, with the low Conscientiousness group having a more peaked distribution compared with the other groups (see Figure 7.3). The dispersion of scores was substantially larger for students in the high Conscientiousness group, suggesting this group was more heterogeneous compared to the other two groups. The median scores for each group were similar, thus suggesting that the level of Conscientiousness may not necessarily determine/be related to test performance, at least for the sample employed. There were several outliers for each group representing students who scored very low in their midterm test.

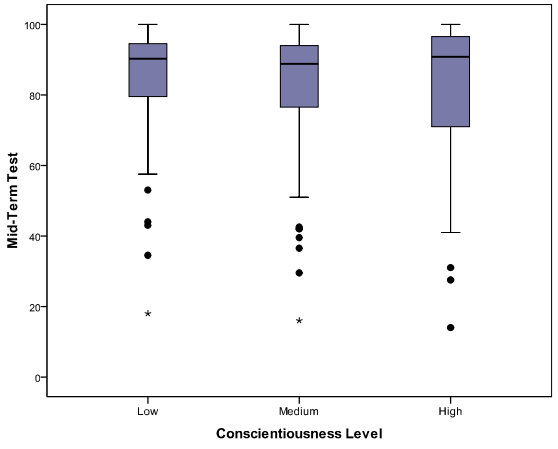


Figure 7.3 Comparison of test scores between groups

Figure 7.4 shows the boxplots of final exam scores for each group of Conscientiousness levels. The distributions of scores between the boxplots have a similar spread and they are all negatively skewed. The median scores and the upper quartiles between groups were also similar, thus suggesting, at least based on the boxplots, that there were no noticeable differences in final exam performance between the different levels of Conscientiousness.

Note that each group had a few outliers, representing students who scored very low in their final exam.

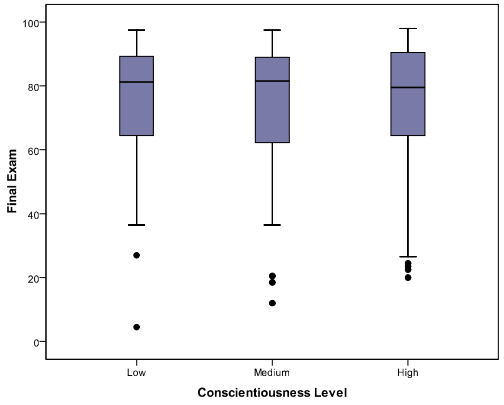


Figure 7.4 Comparison of final exam scores between groups

Table 7.3 shows the results from applying a bivariate correlation test to measure the association between FFM variables and academic performance. Contrary to our expectations, there is no significant relationship between level of Conscientiousness and performance. However, students' performance in assignments, and a final exam showed a significant positive relationship with Openness to experience. The strongest positive correlation was between final exam scores and Openness to experience, $r(210) = 0.20$, $p = 0.007$. The findings regarding Openness to experience were consistent with those from our previous experiment (see Chapter 6).

Table 7.3 Correlation between academic performance and personality factors (N=212)

	Assign	Test	Final	Extrav.	Agreeab.	Consc.	Neuro.
Assign	1						
Test	0.58**	1					
Final	0.66**	0.83**	1				
Extrav.	-0.04	-0.08	-0.09	1			
Agreeab.	0.03	0.06	0.04	-0.08	1		
Consc.	0.02	-0.07	-0.02	0.24**	0.17*	1	
Neuro.	-0.01	-0.06	-0.05	-0.33**	-0.06	-0.34**	1
Openn.	**0.19****	0.12	**0.20****	0.11	0.14**	0.02	-0.11

**. Correlation is significant at the 0.01 level (1-tailed).
*. Correlation is significant at the 0.05 level (1-tailed).

Based on our correlation analysis, we have conducted a supplementary analysis to gain insight into the correlations between the dependant variables and the narrower facets of Conscientiousness. The narrower facets are more behaviorally specific than the broad trait of Conscientiousness, thus it helps in identifying which behavioral tendencies relate to the performance criterion (LePine, Colquitt and Erez, 2000). The six facets of Conscientiousness are *self-efficacy*, *orderliness*, *dutifulness*, *achievement-striving*, *self-discipline*, and *cautiousness* (Costa & McCrae, 1995). Table 7.4 presents the results of the supplementary analysis. These results show that there were no significant correlations between any of the

achievement facets (i.e. achievement striving, self-efficacy, and self-discipline) and the measures of academic performance. There was a negative significant correlation between orderliness and the midterm test, however overall results indicate that none of the Conscientiousness facets strongly positively correlated with students' academic performance. Thus, Conscientiousness appeared to be insignificant in differentiating the academic performance of students who pair programmed, at least in the sample employed in this study.

Table 7.4 Correlations of Conscientiousness facets and academic performance (N=212)

	Assign	Test	Final	Efficacy	Orderliness	Dutifulness	Achievement	Discipline
Assign	1							
Test	0.58**	1						
Final	0.66**	0.83**	1					
Efficacy	0.12	0.01	0.02	1				
Orderliness	-0.04	**-0.18***	-0.12	-0.06	1			
Dutifulness	0.07	0.01	0.01	0.15*	0.15*	1		
Achievement	-0.00	-0.04	-0.05	0.45**	0.15*	0.29**	1	
Discipline	-0.04	-0.09	-0.08	0.39**	0.37**	0.35**	0.48**	1
Cautious.	-0.09	-0.02	-0.02	0.17*	0.23**	0.26**	0.33**	0.37**

**. Correlation is significant at the 0.01 level (1-tailed).
*. Correlation is significant at the 0.05 level (1-tailed).

7.6.3 Hypothesis Testing

The null hypothesis was tested using the one-way analysis of variance (ANOVA) test to analyze whether there was any significant differences in academic performance between the three levels of Conscientiousness (low, medium, and high). The one-way ANOVA procedure was reported to be robust and also a technique that can be relied upon even when distributional assumptions are violated (Morgan et al., 2004).

Table 7.5 Mean and standard deviation of paired students of different level of Conscientiousness

	Conscientiousness Level	N	Mean	SD
Assignments	Low Consc	66	11.42	4.00
	Medium Consc.	76	11.74	4.05
	High Consc.	70	12.11	3.63
	Total	212	11.76	3.89
Test Scores	Low Consc	66	83.92	16.80
	Medium Consc.	76	81.76	19.00
	High Consc.	70	82.05	20.43
	Total	212	82.53	18.78
Final Exam	Low Consc	64	74.95	19.70
	Medium Consc.	75	73.64	20.10
	High Consc.	69	73.32	21.53
	Total	208	73.94	20.38

Table 7.5 provides the mean values and standard deviation values for academic performance for each group. Mean differences are almost similar for all measures of PP's effectiveness (assignments, test, and final exam). The Levene test showed that the variances between groups were not significant; therefore, the assumption for homogeneity of variance was not violated (see Table 7.6). The overall *F* values for the three ANOVA are presented in Table 7.7. The results show that there were no significant differences in academic performance

between the three groups of Conscientiousness (i.e. $F(2, 209)= 0.54$, $p = 0.58$, for assignments; $F(2, 209) = 0.27$, $p = 0.77$, for test; $F(2, 205) = 0.12$, $p = 0.89$, for final exam). Since none of the F values were statistically significant, no post-hoc analysis was needed. Our results indicate that we could not find strong support to reject the null hypothesis. Thus, based on our data, we found that PP's effectiveness was not affected by differences in Conscientiousness levels among paired students.

Table 7.6 Test of Homogeneity of variances

	Levene Statistic	df1	df2	Sig.
Assignments	0.57	2	209	0.57
Test Scores	1.85	2	209	0.16
Final Exam	0.33	2	205	0.72

Table 7.7 ANOVA results

		Sum of Squares	df	Mean Squares	F	Sig.
Assign.	Between Groups	16.47	2	8.23	0.54	0.58
	Within Groups	3181.42	209	15.22		
	Total	3197.89	211			
Test	Between Groups	188.49	2	94.25	0.27	0.77
	Within Groups	74231.11	209	355.17		
	Total	74419.61	211			
Final Exam	Between Groups	99.19	2	49.59	0.12	0.89
	Within Groups	85877.40	205	418.91		
	Total	85976.59	207			

7.6.4 Statistical Power Analysis

We conducted a post hoc power analysis to compute the statistical power of the ANOVA test employed to analyze the data gathered in our experiment. The power analysis was carried out to determine the ability of the statistical test to detect an effect, if the effect truly exists. As mentioned by Cohen (1988), the power indicates the probability of correctly rejecting the null hypothesis. In the event where the null hypothesis is wrongly rejected, the risk of committing this is known as "Type 1 error"; denoted by the significance or alpha (α) level used in the study (Cohen, 1988). Likewise, if the null hypothesis is wrongly accepted, the risk of committing this is known as "Type 2 error"; denoted by the β value. Thus, the power of the test is equivalent with the probability of rejecting the false null hypothesis, (1 – β) (Cohen, 1988).

Due to the null hypothesis investigated herein not being rejected (non-significant results), the post hoc power analysis helps in interpreting the results. Non-significant findings may indicate that the treatment has no effect (i.e. effects of personality trait Conscientiousness in our experiment may not prominent), or that the results are inconclusive if the power is found to be low. The software package G*Power (Version 3.1.2) (Faul et al., 2007) was used to compute the statistical power of our results.

Statistical power is calculated based on the input of three parameters: i) the significance level, α; ii) the sample size; and iii) the effect size of the population (Faul et al., 2007). In G*Power, the computation of the effect size is based on the Cohen's approach (Cohen,

1988). The analysis was carried out separately for each dependent variable using the F-test family of the one-way ANOVA.

Tables 7.8 to 7.10 show the protocols of the power analysis where the input and output parameters were specified. As can be seen from the tables, the effect size and the power of statistical test were considered to be very low. For instance, the power equal to 0.14 indicates that we can only have approximately 14% chance of correctly rejecting the null hypotheses if it is false (see Table 7.8). Similarly, the other analyses showed the power of 0.10 and 0.08 based on ANOVA test for both the midterm test and the final exam, respectively (see Tables 7.9 and 7.10).

The small effect size derived from the analysis also affects the level of statistical power. This is because the ability to detect effects even when they exist is more difficult when the effect size is small (Miller et al., 1997). Note that a post hoc analysis requires the measure of effect size to be based on the *population* effect size. Nevertheless, the true population effect size is never actually known (Abraham & Russel, 2008). Therefore the effect size in this analysis is estimated based on the sample data and thus the effect may not be essentially identical to the effect size in the population.

Table 7.8 Power Analysis Protocol (Assignments)

F tests – ANOVA: Fixed effects, omnibus, one-way		
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size f	= 0.07
	α err prob	= 0.05
	Total sample size	= 212
	Number of groups	= 3
Output:	Noncentrality parameter λ	= 1.10
	Critical F	= 3.04
	Numerator df	= 2
	Denominator df	= 209
	Power (1-β err prob)	= 0.14

Table 7.9 Power Analysis Protocol (Midterm Test)

F tests – ANOVA: Fixed effects, omnibus, one-way		
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size f	= 0.06
	α err prob	= 0.05
	Total sample size	= 212
	Number of groups	= 3
Output:	Noncentrality parameter λ	= 0.64
	Critical F	= 3.04
	Numerator df	= 2
	Denominator df	= 209
	Power (1-β err prob)	= 0.10

Table 7.10 Power Analysis Protocol (Final Exam)

F tests – ANOVA: Fixed effects, omnibus, one-way		
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size f	= 0.04
	α err prob	= 0.05
	Total sample size	= 208
	Number of groups	= 3
Output:	Noncentrality parameter λ	= 0.35
	Critical F	= 3.04
	Numerator df	= 2
	Denominator df	= 205
	Power (1-β err prob)	= 0.08

The plots shown in Figure 7.5 were generated from the power analyses. The graph visualizes the power of the F test for the ANOVA as a function of the sample size. By comparing the slopes of the three curves, the greater the significance level (alpha value), the more sensitive the statistical power to the sample size. In order to achieve the desired high level of statistical power (i.e. 0.80) at the specified parameters (number of groups = 3; effect size = 0.07; alpha = 0.05) would require an extraordinarily large sample size (i.e. over 1400 sample). Given this small effect size, it may not be feasible or practical to obtain such a huge sample size if we were to replicate the study. Note that we followed the classifications of effect size as reported by Dyba et al. (2006).

At a moderate effect size, increasing the sample size in future replications of the study will also increase the statistical power value (see Figure 7.6). Thus, the non-significant results may not hold if a greater sample size is employed in future replication of the study. Given that our analysis presented low statistical power, it would be insufficient to provide confidence that these results correspond to what would be most likely observed when we investigate the effects of Conscientiousness on paired students' academic performance in a higher education environment.

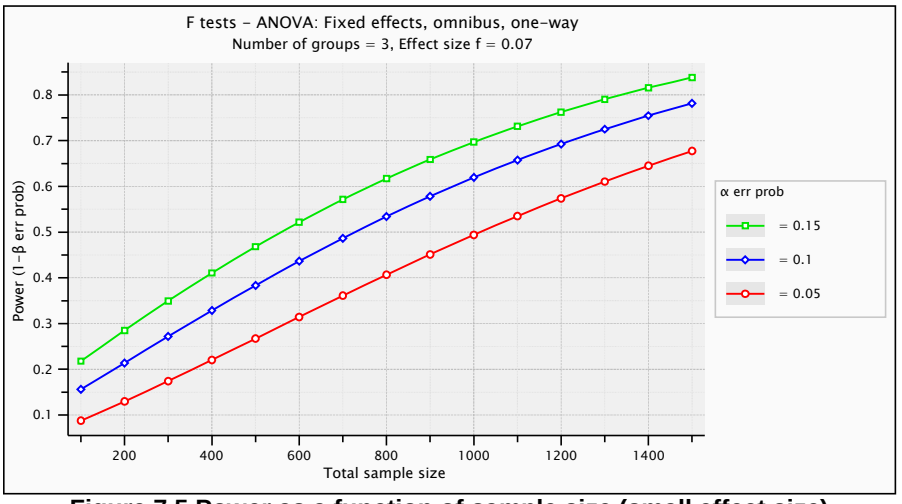


Figure 7.5 Power as a function of sample size (small effect size)

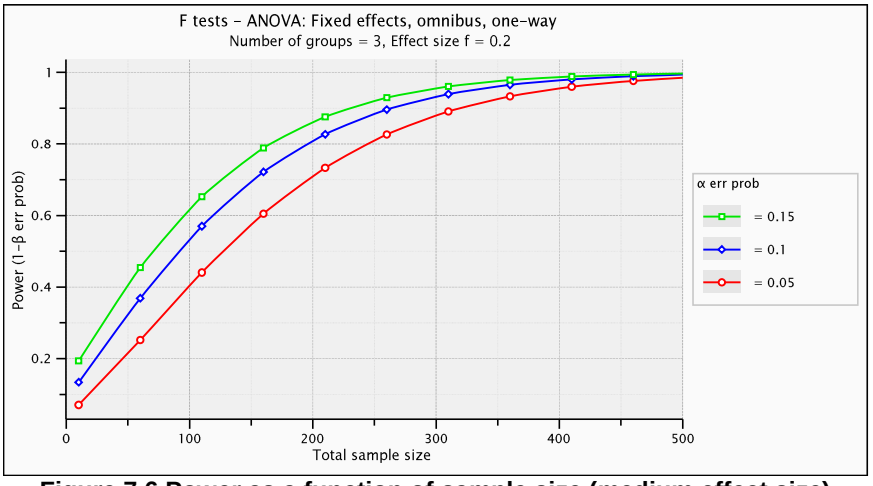


Figure 7.6 Power as a function of sample size (medium effect size)

7.6.5 Results for Satisfaction and Confidence

In order to measure paired students' levels of satisfaction and confidence, questionnaires were used to gather data after each tutorial session. Altogether there were nine weeks of tutorials starting from the second week of the semester until the final week of tutorials. Data were analyzed separately as each tutorial was treated as a single independent "mini-experiment".

The response rate of the post-experimental survey was of approximately 67% during the first week, but decreased to 42% for the last week of tutorials. One of the questions used in the questionnaire asked subjects to "*Please rate how satisfied are you working with your partner?*" and their response based on the five-point Likert scale is shown in Figure 7.7. Our data analysis showed that on average 111 (90.2%) out of an average of 123 students attending the tutorials were satisfied working with their partner (more than 50% reported that they were very satisfied).

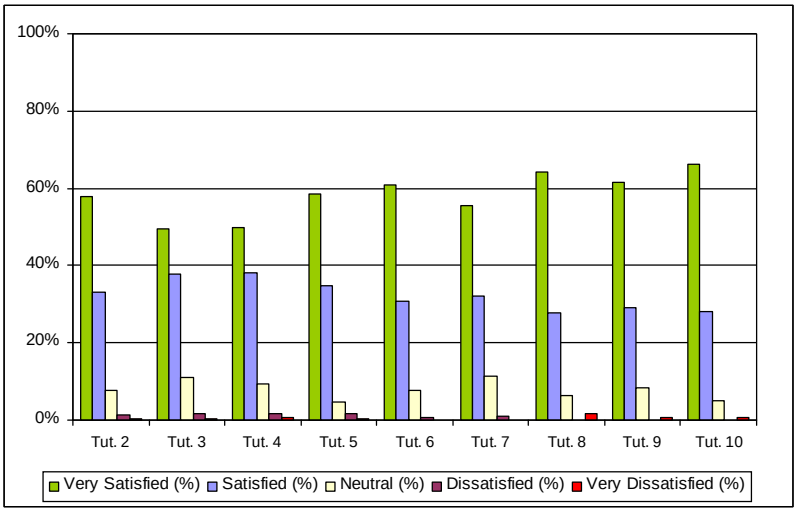


Figure 7.7 Survey on PP satisfaction

The Kruskal-Wallis test was used to compare satisfaction levels between the Conscientiousness groups. This statistical test was chosen because our dependent variable was measured on an ordinal scale. Table 7.11 shows the mean satisfaction rank of paired students. The group with the highest mean rank had the highest level of satisfaction. Although paired students of high Conscientiousness appeared to obtain higher ranks in most of the experiment units, these differences were not always significant. As can be seen in Table 7.11, only Tutorial 2 and 7 showed a significant value ($p = 0.00$ and $p=0.02$ respectively) but overall results demonstrated that the satisfaction level of paired students were not affected by students' levels of Conscientiousness.

Table 7.11 Mean rank for satisfaction level

	Consc. Level	N	Mean Rank	Sig.
Tut. 2 N=147	Low	38	57.87	0.00
	Medium	53	68.73	
	High	56	88.99	
Tut. 3 N=131	Low	27	58.46	0.34
	Medium	57	65.76	
	High	47	70.62	
Tut. 4 N=156	Low	32	68.02	0.06
	Medium	68	75.35	
	High	56	88.32	
Tut. 5 N=132	Low	36	57.25	0.16
	Medium	43	69.12	
	High	53	70.66	
Tut. 6 N=132	Low	26	67.00	0.28
	Medium	39	59.85	
	High	67	70.18	
Tut. 7 N=115	Low	28	59.96	0.02
	Medium	42	48.19	
	High	45	65.93	
Tut. 8 N=106	Low	28	51.52	0.49
	Medium	35	50.73	
	High	43	57.05	
Tut. 9 N=94	Low	18	46.11	0.44
	Medium	34	44.01	
	High	42	50.92	
Tut.10 N=91	Low	16	49.75	0.76
	Medium	37	45.55	

	High	38	44.86	

Data on students' levels of confidence working in pairs were also gathered using the questionnaire. When asked "*How do you rate your level of confidence solving the exercises with your partner*", an average of 107 students (87.7%) who attended the tutorial classes answered that their level of confidence solving tasks with their partner was high when working in pairs (see Figure 7.8). Table 7.12 shows the mean rank of the Kruskal-Wallis test for the paired students' confidence level. Although paired students of high Conscientiousness appeared to score higher ranks in most of the experiment units, there were only three tutorials that presented a significance difference in confidence level based on the Conscientiousness levels: Tutorial 2 (X^2 (2,147)=9.51, ρ = 0.01), Tutorial 6 (X^2 (2,132)=6.23, ρ = 0.04), and Tutorial 7 (X^2 (2,115)=8.47, ρ = 0.01). In these tutorials, paired students of high Conscientiousness level indicated a greater satisfaction level compared with the other groups.

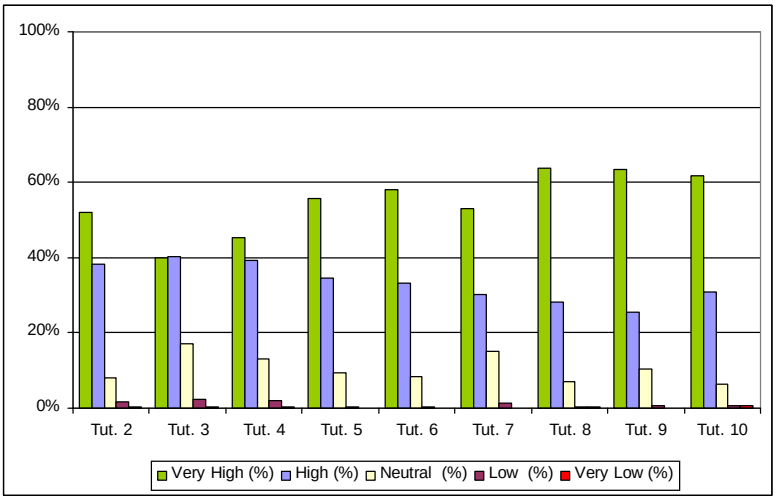


Figure 7.8 Survey on PP confidence

Table 7.12 Mean rank for confidence level

	Consc. Level	N	Mean Rank	Sig.
Tut. 2 N=147	Low	38	62.38	0.01
	Medium	53	69.72	
	High	56	85.94	
Tut. 3 N=131	Low	27	64.56	0.87
	Medium	57	64.86	
	High	47	68.21	
Tut. 4 N=156	Low	32	71.11	0.17
	Medium	68	75.21	
	High	56	86.71	
Tut. 5 N=132	Low	36	60.85	0.45
	Medium	43	66.66	
	High	53	70.21	
Tut. 6 N=132	Low	26	70.13	0.04
	Medium	39	55.47	
	High	67	71.51	
Tut. 7 N=115	Low	28	59.75	0.01
	Medium	42	48.08	
	High	45	66.17	
Tut. 8 N=105	Low	27	50.48	0.63
	Medium	35	51.46	
	High	43	55.84	

Tut. 9 N=94	Low Medium High	18 34 42	45.81 44.29 50.82	0.45
Tut. 10 N=90	Low Medium High	15 37 38	46.37 45.53 45.13	0.98

In addition to measuring the satisfaction and confidence level, students' feedback on their pairing experience were also gathered. The questions asked were:

"I felt that working with this partner was a productive experience." (Q1)

"I enjoyed working with my partner." (Q2)

"My motivation level increased when working with my partner." (Q3)

In our analysis, on average 118 out of 124 students (95%) responded that their experience working with the partner was a helpful experience (see Figure 7.9). In their written feedback, students commented that PP is a good way to learn and help them understand the topic better. In terms of enjoyment (Q2), most students (117 out of 124, 94%) agreed that working in pairs was an enjoyable experience (see Figure 7.10). Similarly, students responded that their motivation levels increased when working with their partner. On average 107 out of 124 students (86%) agreed with statement in Q3 (see Figure 7.11).

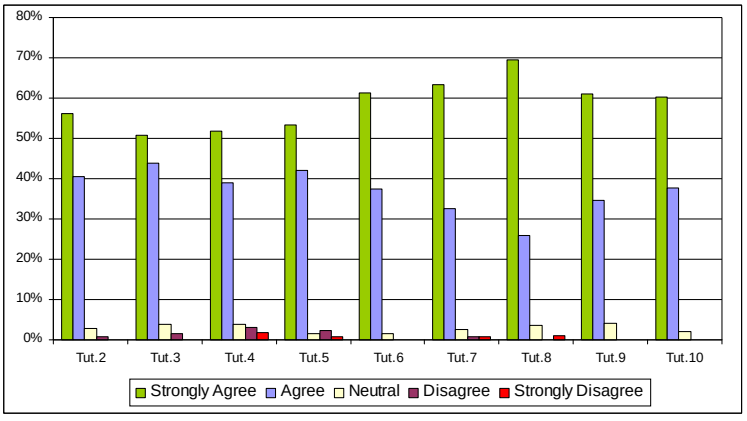


Figure 7.9 PP as a productive experience (Q1)

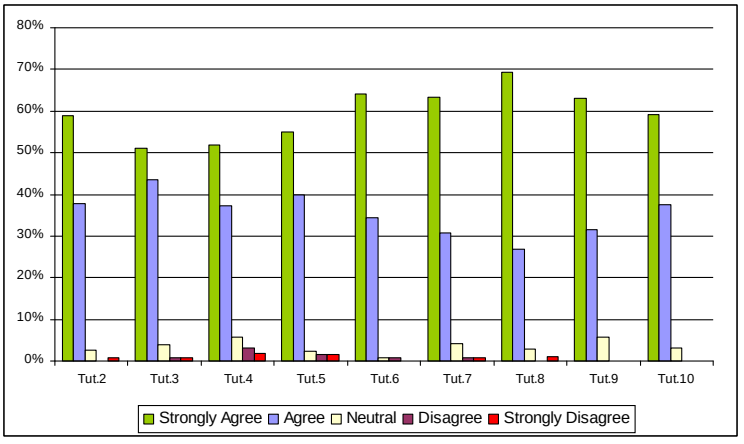


Figure 7.10 Enjoyment (Q2)

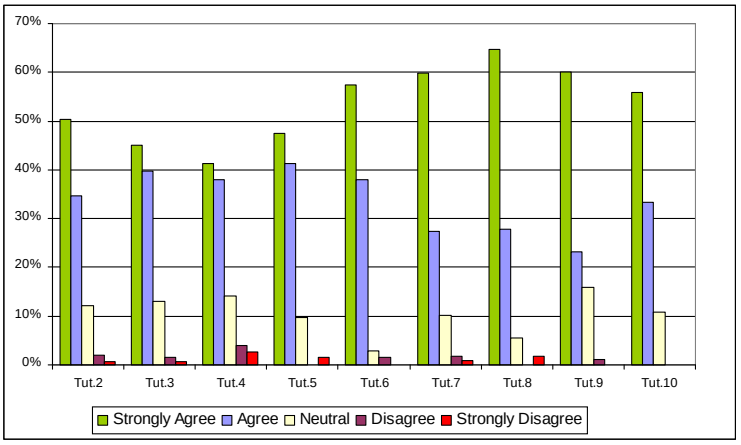


Figure 7.11 Motivation level (Q3)

7.7 Discussion

The focus of this experiment was to investigate the effect that level of Conscientiousness may have on PP's effectiveness as a pedagogical tool. Our results showed that the academic performance of paired students was not significantly affected by their level of Conscientiousness. Although these results seemed to contradict some of the previous findings reported in the literature (Zyphur et al., 2008; Busato et al., 2000; Chamorro-Premuzic & Furnham, 2003b), there is also evidence showing that level of Conscientiousness may not be always prominent in affecting the performance of student teams (Mohammed, Mathieu, & Bartlett, 2002; Kichuk & Wiesner, 1997; Peeters et al., 2006).

Peeters et al. (2006) argued that the effect of Conscientiousness can be absent in student teams due to the short period of time available for teams to complete a task and also due to the low levels of interdependency among team members. As our students practiced PP for less than two hours only once a week during the tutorials, this might explain why we did not find a significant difference between the three different types of Conscientiousness levels and performance.

An investigation into self-managed groups also did not find a significant relationship between Conscientiousness and team outcome (Barry & Stewart, 1997). They suggested that "C*onscientiousness may become less important in team-based tasks because groups are able to recognize and compensate for the lack of conscientious individuals*" (p. 76, Barry & Stewart, 1997). In another study, LePine et al. (2000) showed that there were differential effects for facets of Conscientiousness in teams' decision making performance. In their study, the effects of Conscientiousness were found to be the opposite of what they expected, and their further analysis showed that the findings were due to the traits reflecting a dependability facet rather than an achievement facet. In our supplementary analyses, we found that none of the Conscientiousness facets were positively correlated with our dependant variables. Thus, based on our data we found that the effect of Conscientiousness on PP's effectiveness appeared to be insignificant.

Another possible explanation for our findings relates to the fact that the assessment of tasks/assignments contributed towards a student's course grade, therefore students may have tended to perform well regardless of their personality attributes. As reported by Kichuk & Wiesner (1997), "*The relative novelty of being a university student and the perceived consequences of not performing well may have caused most students to behave conscientiously while doing the task regardless of how they scored on the personality profile*" (p. 211). This can also explain why team composition based on personality traits differed between academic (i.e. lab setting) and industry environments (i.e. actual field setting) in relation to team performance, as reported by Bell (2007). The small effect size obtained from our experiment may indicate that the effects of the personality trait Conscientiousness on paired students' academic performance may be trivial. The non-significant findings from this experiment resulted from the low level of statistical power generated in the power analysis.

Of all the FFM personality trait measures, our results showed that Openness to experience showed the most prominent relationship with PP's effectiveness (measured by academic performance). These findings were consistent with those found in our previous experiment (see Chapter 6), and also corroborate results from other studies (Farsides & Woodfield, 2003; Chamorro-Premuzic & Furnham, 2008; Fruyt & Mervielde, 1996).

Previous research (Blickle, 1996; Staudinger, Maciel, Smith, & Baltes, 1998) suggests that Openness to experience facilitates the use of learning strategies, and those students with a relatively high level of Openness to experience were described as being foresighted, intelligent, and resourceful (De Raad & Schouwenburg, 1996). It was also expected that Openness to experience to be more influential on performance for the tasks that require creativity or tackling of abstract problems (Driskell et al., 1987). Within the context of our study, paired students were given programming exercises typically considered as abstract by their nature. Therefore, this may explain why in our case Openness to experience may have influenced PP's effectiveness far more than the Conscientiousness aspect. In Chapter 10, we investigate the effects of Openness to experience on the PP's effectiveness.

7.8 Threats to the Validity

There are several potential threats to the validity of our results. The first relates to the sample size (212 students). A larger sample size would have helped statistical power of the results (Miller et al., 1997; Cohen, 1988). This is because the probability of detecting the phenomenon is greater when increasing the sample size (Cohen, 1988).

Another limitation relates to the construct validity of our dependent variable. Herein we used students' academic performance as a surrogate measure of PP's effectiveness. However students' performance may also be affected by their cognitive or mental ability. Regardless of one's personality behavior while pairing, students may perform well due to their ability or competency in programming. However, since the aim of this experiment is to improve students' learning due to practicing PP throughout the entire semester, measuring their academic performance is in our view appropriate for use in our context. Note that with

this set of instruments we are unable to measure the amount learned as we didn't analyze before/after PP competency.

Another issue relates to the handling of the lectures and tutorials. The course was taught by three different instructors throughout the semester, each handling a three weeks block of teaching. Similarly, the weekly tutorials were also run by three different tutors. In these circumstances, differences in teaching style may have had influence students' motivation and their comprehension level of the course.

A further limitation is that our experiment did not control for the effects of gender. Although earlier meta-analysis suggested that gender may affect personality traits (Feingold, 1994), secondary analyses of personality data based on FFM reported that the difference was generally subtle or small relative to individual personality variation within a single gender group (Costa, Terracciano, & McCrae, 2001). In particular, gender differences only appeared pervasive in facets of Neuroticism and Agreeableness, and fairly negligible for Conscientiousness (Costa et al., 2001). Therefore, this issue might not play a significant role in the results obtained in our experiment.

7.9 Summary

In summary, our experiment did not reject the null hypotheses, thus it did not provide any evidence for distinguishing the performance of paired students between different levels of Conscientiousness. The non-significance of our findings is probably related to the inadequate statistical power generated from the statistical tests. Despite the counterintuitive findings regarding the effects of Conscientiousness, the results of the formal experiment showed a positive correlation between Openness to experience and measures of PP's effectiveness in assignments, and final exam. This corroborates findings of existing studies reported by Chamorro-Premuzic & Furnham (2008), Farsides & Woodfield (2003) and also our previous study (Salleh et al., 2009).

On average, 90% of students were satisfied with the PP experience. Similarly, 88% of students responded that their confidence level increased when working in pairs. This evidence suggests that regardless of the variation in students' Conscientiousness level, PP not only caused an increase of satisfaction and confidence level, but also brought enjoyment to the class and enhanced students' learning motivation. The current findings add to our understanding of the effect of Conscientiousness towards students' academic performance when practicing PP in an introductory programming course. In the next chapter, the effects of Conscientiousness are investigated in a more advanced CS course.

Chapter 8

THE THIRD EXPERIMENT

This chapter describes the third experiment conducted during the first semester of 2009 at the University of Auckland. The subjects who participated in the experiment were second year undergraduate students enrolled in the Software Design and Construction course (COMPSCI 230). The objectives and details of this experiment are explained in the following sections, followed by a discussion of the results and the threats to the validity of our findings.

8.1 Experimental Objectives

The formal experiment described herein was conducted during the same academic semester as the experiment reported in Chapter 7 (2nd experiment). In that experiment we investigated whether or not the differences in levels of Conscientiousness (i.e. low/medium/high) among paired undergraduate students affected their academic performance. In the present experiment, we also focused on the Conscientiousness factor of the FFM, investigating the same hypothesis investigated in the 2nd experiment. However in this experiment we investigated the effects of Conscientiousness among more mature students involved in a more advanced computing course. Conscientiousness was chosen, and also used in the present experiment, because this factor was reported to be strongly associated with team performance as well as academic success (Chamorro-premuzic & Furnham, 2003b; Fruyt & Mervielde, 1996; Duff et al., 2004).

The results from our 2nd experiment did not provide strong support to distinguish performance between paired students of different levels of Conscientiousness for introductory programming tasks. However these findings cannot be generalized to a wider population due to a lack of statistical power. The present experiment (3rd experiment) was therefore used to investigate whether similar results to those obtained in the 2nd experiment appear when applying the same experimental setup to a different group of subjects and programming tasks.

The purpose of this experiment was to investigate if the effectiveness of PP as a pedagogical tool for CS/SE education could be improved by investigating the influence of the Conscientiousness factor of the FFM personality model, towards a pair's academic performance. The main objectives of the experiment are to increase students' satisfaction and the amount of students' learning. These outcomes are reflected in their academic performance, shown by the students' performance in assignments, a midterm test and the final exam. These research objectives were outlined using the Goal/Question/Metric (GQM) framework (Basili, Shull, and Lanubile, 1999) shown in Table 8.1. The detailed goal definition for the experiment is indicated as follows:

Object of study: PP technique.

Purpose: To improve the effectiveness of PP as a pedagogical tool in higher education institutions.

Focus: To investigate the influence of Conscientiousness factor of FFM personality model that can potentially affect the success of the PP practice in a more advance CS/SE course.

Perspective: From the point of view of the researcher.

Context: In the context of undergraduate CS/SE students.

Table 8.1 GQM definition

Goal(s)	Question(s)	Metric(s)
To investigate the effect of Conscientiousness towards successful pair configuration in software design course.	Do differences in Conscientiousness level within a pair affect the pair's academic performance?	Students' academic achievement measured by assignments, midterm test and final exam scores.
To investigate the level of satisfaction of paired CS students.	Were students feeling satisfied working in pairs?	PP questionnaire on satisfaction level.
To investigate the level of confidence of paired CS students.	Were students feeling confident working in pairs?	PP questionnaire on confidence level.

8.2 Experimental Context

Our 3rd experiment was conducted during the first semester of 2009 and involved second year undergraduate students enrolled in a Software Design and Construction course (COMPSCI 230). The teaching of this course consisted of ten weeks of lectures (30 lectures, each lasting for an hour) and ten weeks of tutorials (each lasting for an hour), none of which requiring compulsory attendance (however attendance is expected). The tutorials were held in a computer lab run by a tutor. During the tutorials, students who participated in the experiment worked in pairs when solving the exercises. Each of the tutorial sessions was treated as an independent experiment thus students' feedback regarding the pairing experience was gathered every session. In this course, students learnt about the concept of software application development in Java, including the GUI framework, multithreading, and some aspects regarding software quality. They were required to submit two individual assignments on object-oriented software development and concurrent programming. Students willing to participate in the experiments were required to sign a consent form as to fulfill the ethical requirements of the University of Auckland's Human Participant Ethics Committee.

8.3 Hypothesis

Of the five personality constructs based on the FFM, Conscientiousness has been found to have a positive association with achievement and competency (McCrae & John, 1992; Busato et al., 2000). Some of the positive attributes of highly conscientious individuals include being hard-working, reliable, organized, purposeful, and diligent. In contrast, those who are less

conscientious typically possess the opposite attributes such as being disorganized, irresponsible, and impulsive (Driskell et al., 2006).

It was reported that a team consisting of a higher composite level of Conscientiousness positively influences knowledge sharing and team performance (Hsu, Wu, & Yeh., 2007; Barrick et al., 1998). Busato et al. (2000) report Conscientiousness as a consistent and positive predictor of academic success. Other findings supporting the positive association between Conscientiousness and academic performance were reported by Chamorro-Premuzic & Furnham (2003b); Furnham, Chamorro-Premuzic & McDougall (2003); and Dollinger & Orf (1991). Such a positive association is primarily due to two important facets of Conscientiousness, known as dutifulness and achievement striving (Chamorro-Premuzic & Furnham, 2003b). Therefore, we posited that Conscientiousness may influence the academic performance of students practicing PP. In this experiment, we investigated the following hypothesis:

H_O: Differences in Conscientiousness level do not affect the academic performance of students who pair programmed.

which is contrasted by the following alternative hypothesis:

H_A: Differences in Conscientiousness level affect the academic performance of students who pair programmed.

Table 8.2 shows the categorization of pairs according to students' level of Conscientiousness. Pair (C_{High}, C_{High}) denotes a combination where both students have high levels of Conscientiousness. This experiment compared the performance of students in these groups based on their academic achievement in the course. Our experiment also looked into the association between an individual's personality score with their academic performance, and level of satisfaction and confidence when working in pairs.

Table 8.2 Pair configuration

Conscientiousness level	Pairing groups
High	Pair (C_{High}, C_{High})
Medium	Pair (C_{Med}, C_{Med})
Low	Pair (C_{Low}, C_{Low})

8.4 Variables

Our SLR's synthesis of evidence showed that measuring PP's effectiveness could be achieved using "academic performance", "technical productivity", "program quality", or "satisfaction" (Salleh et al., 2010). Since our experiment aimed at facilitating CS/SE students through the practice of PP, the metrics selected to measure PP's effectiveness were "academic performance" and students' "satisfaction". Hence, PP's effectiveness and satisfaction were our dependent variables and level of Conscientiousness our independent variable (single-factor).

In this experiment, PP's effectiveness was measured using assignments (20%), a midterm test (15%) and final exam scores (65%). Level of satisfaction was measured using a

questionnaire where all questions employed a five-point Likert-scale. We employed the same set of instruments as in the 2nd experiment (see Chapter 7).

8.5 Experimental Procedure

The experiment took place during the tutorial sessions offered as part of the COMPSCI 230 course. During the first week of the semester, students' personality data were gathered using the online IPIP-NEO test. Students who signed the consent form were required to fill out the online personality test using the IPIP-NEO inventory.

We gathered the results of the personality test in order to allocate partners. For this purpose, the numerical personality scores relating to the Conscientiousness factor were used to assign students to one of three possible groups: low, medium or higher levels of Conscientiousness. Based on the distribution of scores for the Conscientiousness trait, the grouping or the classification of Conscientiousness levels used in this experiment was based on this scores' range: lowest 40% (low Conscientiousness); middle 30% (medium Conscientiousness); and the highest 30% (high Conscientiousness)

The allocation of pairs within each group was done randomly. Since we only had Conscientiousness as our independent variable, our hypothesis was investigated using a "*single factor between-group design*" as our experimental design (Morgan et al., 2004).

Each tutorial lasted for only one hour. During this time, the tutor explained the topic for about 15 minutes, and the students completed exercises for the remaining 45 minutes. Due to the limited time, the swapping of roles (between driver and the navigator) took place only once. Before the end of every tutorial, students provided feedback relating to working with the partner by filling out a questionnaire (see Appendix B.4). The exercises given during the tutorials were not graded, and thus did not contribute towards students' final grade. In addition, the assignments and a midterm test were graded and completed individually. Students' grades on this course were evaluated based on their achievement in two assignments, a midterm test, and a final exam.

Since tutorial exercises varied from week to week, the experiments were designed in such a way to minimize the confounding factor which might occur due to differences in tasks and exercises' levels of complexity. Therefore, the same set of exercises was given throughout a week.

8.6 Results and Analysis

This section describes the results of the experiment including the demographics relating to the sample population used. The interpretation of results is presented under the discussion section and finally the potential threats to validity of the results are also discussed.

8.6.1 Demographics

A total of 179 students were enrolled in the COMPSCI 230 course during the first semester of 2009. Of these, 147 were male students (82.1%) and 32 were females (17.9%); subjects' age

ranged from 19 to 32 years old (median age = 21 years). Of the 104 students who answered the demographics questionnaire, 75 students (72%) did not have any work experience; however 25 students (24%) indicated that their programming competency was above average.

These demographics also showed that the subjects came from various ethnic backgrounds; the majority being Chinese (46 students, 25.7%), followed by NZ/Pakeha (30 students, 16.8%). Other ethnic groups included South Korean (8 students, 4.5%), Indian (11 students, 6.1%), Asian (6 students, 3.4%), Middle Eastern (3 students, 1.7%), and African (2 students, 1.2%). Of the 179 students, only 79 students (44%) completed the personality test. Of these 179 students, only 77 remained enrolled throughout the semester and sat the midterm test and the final exam. Therefore the sample size used in our analysis was 77.

8.6.2 Data Distribution

The distribution of personality scores based on the FFM test is shown in Figure 8.1. The box in the boxplot represents 50% of the cases, with the upper and the lower whiskers representing the third and first quartiles, respectively. Note that in all boxplots (Figure 8.2 to Figure 8.4), the maximum possible scores for the assignments, test, and final exam are 20, 15, and 65 respectively.

As can be seen in Figure 8.1, Conscientiousness presented a more heterogeneous distribution followed by Agreeableness, Extraversion, and Neuroticism. The median value for Agreeableness was the highest among all factors whereas the lowest median belongs to the Openness to experience. Distribution of scores for Openness to experience is also positively skewed. The additional circle outside the range of distribution is an outlier. Within this context, it represents students who obtained a very high score on the Openness to experience and Neuroticism.

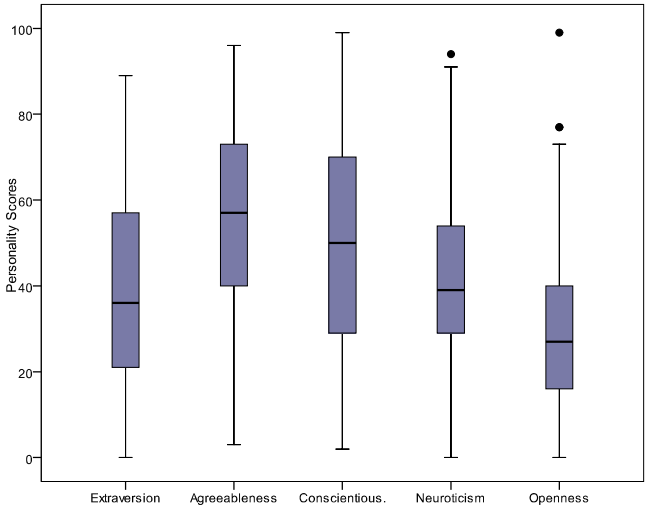


Figure 8.1 Comparison of FFM scores

The boxplots in Figure 8.2 show the distribution of assignment scores for each of the Conscientiousness levels. The flattest distribution belongs to the high Conscientiousness group whereas a more peaked distribution belongs to the medium Conscientiousness group. Both low and high Conscientiousness groups have a negatively skewed distribution. The median for the low Conscientiousness group was slightly higher than the other two. The outliers for the low and high Conscientiousness groups represent the students who obtained extremely low scores for their assignments.

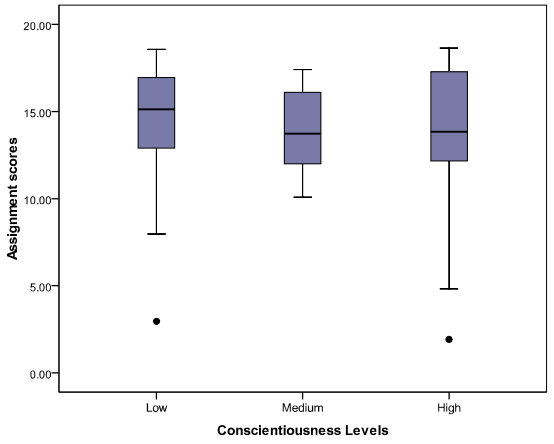


Figure 8.2 Comparison of assignments scores between groups

The boxplots relating to the scores for the midterm test organised by Conscientiousness levels (see Figure 8.3), show a flatter distribution for the low Conscientiousness group, and a more peaked distribution for medium Conscientiousness group. The boxplots also show some variation across groups, with the highest test score belonging to students who were highly conscientious. The median test scores differed widely amongst different Conscientiousness levels. The medium Conscientiousness group had the lowest median score compared with the other groups. The outliers above the third quartile represent cases where students obtained extremely high scores and the outliers below the first quartile indicate cases where the test scores were extremely low.

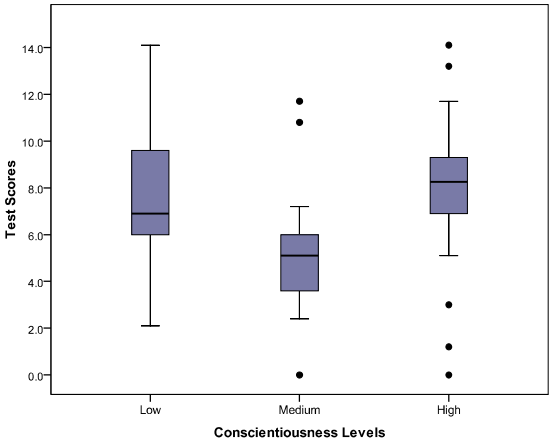


Figure 8.3 Comparison of midterm test scores between groups

Regarding the final exam scores, the medians for the low and high Conscientiousness groups were similar and slightly higher than the median for the medium Conscientiousness group (see Figure 8.4). The high Conscientiousness group obtained the flattest distribution compared with the other two groups, suggesting greater variation of scores. Outliers were present for the low Conscientiousness group.

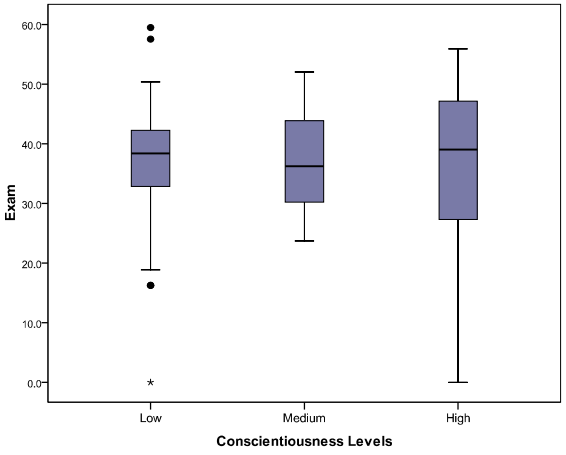


Figure 8.4 Comparison final exam scores between groups

8.6.3 Correlation Analysis

Table 8.3 shows the results of applying the bivariate Pearson's correlation test to measure the association between FFM variables and academic performance. Similar to our findings in the second experiment (see Chapter 7), results showed no significant correlation between paired students' Conscientiousness levels and academic performance. However, students' performance (both in the midterm test, and final exam scores) showed a significant positive correlation with Openness to experience. The strongest statistically significant correlation (positive) was found between final exam scores and Openness to experience, $r(77) = 0.26$, $p = 0.025$, followed by another statistically significant correlation (positive) between the midterm test and Openness to experience, $r(77) = 0.25$, $p = 0.028$. The findings regarding Openness to experience were consistent with those from our 2nd experiment (see Chapter 7).

Table 8.3 Correlation between academic performance and personality factors (N=77)

	Assign	Test	Final	Extrav.	Agreeab.	Consc.	Neuro.
Assign	1						
Test	0.33**	1					
Final	0.61**	0.53**	1				
Extrav.	0.15	0.16	0.12	1			
Agreeab.	-0.11	0.09	0.21	0.15	1		
Consc.	0.00	0.14	0.09	0.18	0.37**	1	
Neuro.	-0.01	-0.10	-0.15	-0.39**	-0.30**	-0.47**	1

Openn.	-0.02	0.25*	0.26*	0.31**	0.39**	0.18	-0.21

**. Correlation is significant at the 0.01 level (1-tailed).
*. Correlation is significant at the 0.05 level (1-tailed).

8.6.4 Hypothesis Testing

In this experiment, the hypothesis focused on investigating whether there is any difference in academic performance amongst paired students of different Conscientiousness levels. The hypothesis was tested using the one-way between-group analysis of variance (ANOVA) (α = 0.05) in order to compare the three levels of Conscientiousness (low, medium, and high) on paired students' academic performance. Table 8.4 shows the mean and standard deviation values for each group. Mean differences are very similar for assignments and final exam scores, whereas means for the midterm test scores varied between Conscientiousness levels.

Table 8.4 Mean and standard deviation of paired students of different level of Conscientiousness

	Conscientiousness Level	N	Mean	SD
Assignments	Low Consc	25	14.19	3.87
	Medium Consc.	31	13.53	3.22
	High Consc.	21	14.37	3.61
	Total	77	13.97	3.52
Test Scores	Low Consc	25	7.57	3.17
	Medium Consc.	31	5.65	3.03
	High Consc.	21	8.40	2.62
	Total	77	7.02	3.16
Final Exam	Low Consc	25	36.53	13.47
	Medium Consc.	31	35.85	8.90
	High Consc.	21	38.78	14.14
	Total	77	36.87	11.95

The results from the Levene test for homogeneity of variances (see Table 8.5) showed that the assumption of homogeneity of data was not violated (F = 0.38, p = 0.69 for assignments; F = 0.94, p = 0.39 for test; and F = 1.11, p = 0.34 for final exam). This test assesses whether the population variances for each of the Conscientiousness groups are significantly different from each other (Carver & Nash, 2006).

Table 8.5 Test of Homogeneity of variances

	Levene Statistic	df1	df2	Sig.
Assignments	0.38	2	74	0.69
Test Scores	0.94	2	74	0.39
Final Exam	1.11	2	74	0.34

The overall F values for the three ANOVA tests are presented in Table 8.6. Based on the ANOVA results (see Table 8.6), we found no significant differences between students' performance (assignments and final exam) and the three groups of Conscientiousness (i.e. $F(2, 74) = 0.43$, p = 0.66, for assignments; $F(2, 74) = 0.38$, p = 0.68, for final exam). However, there was a statistically significant difference in performance between groups of Conscientiousness levels based on the midterm test scores ($F(2, 74) = 5.98$, p = 0.01). A post

hoc comparison was carried out to determine which specific pairs of the test scores' means were significantly different (see Table 8.7).

Table 8.6 ANOVA results

		Sum of Squares	df	Mean Squares	F	Sig.
Assign.	Between Groups	10.74	2	5.37	0.43	0.66
	Within Groups	932.75	74	12.62		
	Total	943.49	76			
Test	Between Groups	105.65	2	52.83	5.98	0.01
	Within Groups	653.65	74	8.83		
	Total	759.30	76			
Final Exam	Between Groups	111.71	2	55.86	0.38	0.68
	Within Groups	10736.09	74	145.09		
	Total	10847.80	76			

Post-hoc comparisons using the Tukey HSD test indicated that the test score's mean score for high Conscientiousness group (M=8.4; SD=2.62) was significantly different from the test score's mean for medium Conscientiousness group (M=5.65; SD=3.03). The result also showed that the low Conscientiousness group and medium Conscientiousness group differed significantly in their midterm test (see Table 8.7). Thus, based on the sample data employed in this experiment, our results indicated that there were no significant differences between academic performance and the three Conscientiousness levels, providing support for the null hypothesis. The exception was on the midterm test, where significant differences were found and the null hypothesis rejected.

Table 8.7 Post hoc test (multiple comparison using Tukey procedure)

Dependent Variables	**(A) Group**	**(B) Group**	**Mean Difference (A-B)**	**Std. Error**	**Sig.**	**95% Confidence Interval**	
						Lower Bound	Upper Bound
Mid-Term Test	Low	Medium	1.92*	0.79	**0.05**	0.01	3.83
		High	-0.83	0.89	0.62	-2.93	1.28
	Medium	Low	-1.92*	0.83	0.17	-3.51	0.48
		High	-2.75*	0.84	**0.05**	-4.76	-0.74
	High	Low	0.83	0.88	0.62	-1.28	2.93
		Medium	2.75*	0.84	**0.05**	0.74	4.78

8.6.5 Statistical Power Analysis

Due to the non-significance of our findings we conducted a post hoc power analysis to compute the statistical power of the ANOVA test employed in our experiment. A statistical power analysis helps to identify whether our findings are likely to also scale up to the entire population under focus, which would indicate that the treatment (Conscientiousness levels) had indeed no effect on performance. If the power is found to be low this indicates that the results are inconclusive. The statistical power analysis presented herein was conducted using G*Power (Version 3.1.2) (Faul et al., 2007).

The computation of statistical power is based on the input of three parameters: i) the significance level, α; ii) the sample size; and iii) an effect size of the population (Faul et al., 2007). In terms of measuring the effect size, G*Power applied the Cohen's approach

(Erdfelder, Faul, & Buchner, 1996; Cohen, 1988). The analysis was carried out separately for each dependent variable using the F-test family of the one-way ANOVA. Table 8.8 to Table 8.10 show the protocols of the power analysis where the input and output parameters were specified. As can be seen, the effect size and the power of the statistical test varied according to the observed dependent variables. Low statistical power was observed for the dependent variables assignment and final exam (0.12 and 0.11, respectively); whereas the observed power was found higher (0.71) for the ANOVA test on the midterm test.

The small effect size derived from the analysis also affects the level of statistical power. This is because the ability to detect effects even when they exist is more difficult when the treatments have very small effect size (Miller et al., 1997; Murphy & Myors, 2003). The high level of power demonstrated on the midterm ANOVA test was due to the medium effect size generated from the power analysis (effect size = 0.32, see Table 8.9). Note that the effect size estimated in this analysis is based on the sample data and thus the effect may not represent the true effect size that exists in the population. This is because the exact true population effect size is generally unknown and should be estimated based upon sample data (Abraham & Russell, 2008; Yuan & Maxwell, 2005).

Table 8.8 Power analysis protocol (assignments)

F tests – ANOVA: Fixed effects, omnibus, one-way		
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size f	= 0.11
	α err prob	= 0.05
	Total sample size	= 77
	Number of groups	= 3
Output:	Noncentrality parameter λ	= 0.86
	Critical F	= 3.12
	Numerator df	= 2
	Denominator df	= 74
	Power (1-β err prob)	= 0.12

Table 8.9 Power analysis Protocol (midterm test)

F tests – ANOVA: Fixed effects, omnibus, one-way		
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size f	= 0.32
	α err prob	= 0.05
	Total sample size	= 77
	Number of groups	= 3
Output:	Noncentrality parameter λ	= 8.23
	Critical F	= 3.12
	Numerator df	= 2
	Denominator df	= 74
	Power (1-β err prob)	= 0.71

Table 8.10 Power analysis protocol (final exam)

F tests – ANOVA: Fixed effects, omnibus, one-way			
Analysis:	Post hoc: Compute achieved power		
Input:	Effect size f	=	0.10
	α err prob	=	0.05
	Total sample size	=	77
	Number of groups	=	3
Output:	Noncentrality parameter λ	=	0.79
	Critical F	=	3.12
	Numerator df	=	2
	Denominator df	=	74
	Power (1-β err prob)	=	0.11

Figure 8.5 shows the graphs of power as a function of the effect size *f* for the three different significance levels (α=0.05; α=0.1; α=0.15). These plots visualize the statistical power value that it would have been likely to detect given the specific sample size and number of groups employed in our experiment. In order to achieve the desired high level of statistical power (i.e. 0.80) at the specified parameters (number of groups = 3; sample size = 77) the effect size should be at least of a medium size (i.e. greater than 0.35) for a significance level of 0.05. Similarly, a medium effect size is required to yield a high level of power (i.e. 80%) with an increase in alpha level (i.e. α ≥ 0.1). We followed the classifications of effect size as reported by Dyba et al. (2006).

The plots in Figure 8.6 and Figure 8.7 showed how the statistical power varies as a function of sample size. When there is a medium effect size (e.g. 0.32), increasing the sample size to approximately 100 subjects will increase the statistical power value to the recommended power of 0.80 (see Figure 8.6). Meanwhile, when the effect size is small (see Figure 8.7) increasing the sample size will also increase the power value but this would incur a higher cost (i.e. in order to get more samples). The results of this analysis indicate that a greater statistical power can be expected if using a larger sample size.

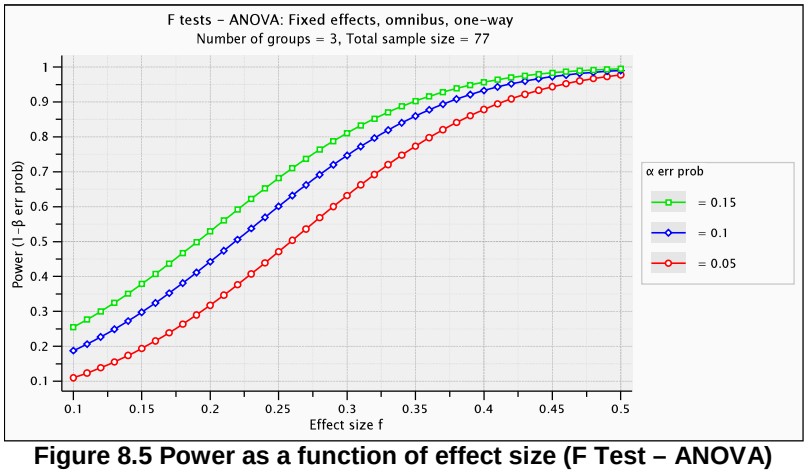


Figure 8.5 Power as a function of effect size (F Test – ANOVA)

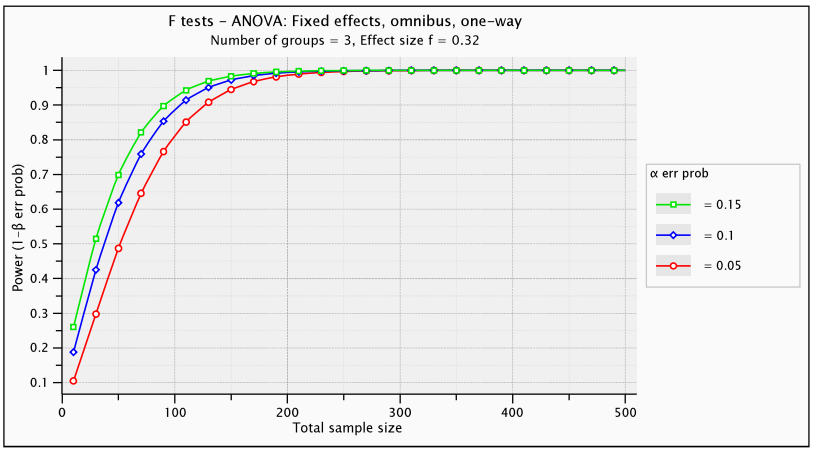


Figure 8.6 Power as a function of sample size (medium effect size)

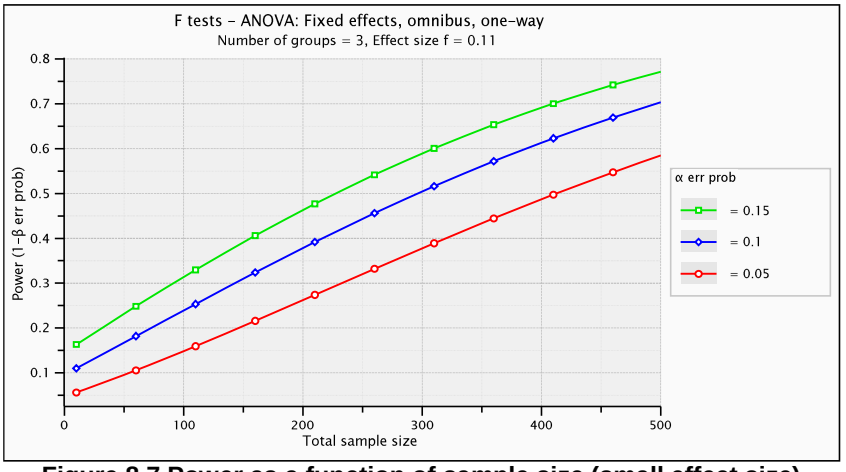


Figure 8.7 Power as a function of sample size (small effect size)

8.6.6 Results for Satisfaction and Confidence

At the end of each tutorial session, students were given a questionnaire to fill out (Appendix B.4) relating to their satisfaction and confidence working with their partner. Altogether there were nine weeks of tutorials starting from the second week of the semester until the final week of tutorials.

The response rate of the post-experimental survey was approximately 77% during the first week, but decreased to 30% for the last week of tutorials. One of the questions asked was "*Please rate how satisfied are you working with your partner?*" and their response based on the five-point Likert scale is shown in Figure 8.8. Our data analysis showed that on average 22 (79%), out of an average of 28 students attending the tutorials, were satisfied working with their partner (more than 38% reported that they were very satisfied).

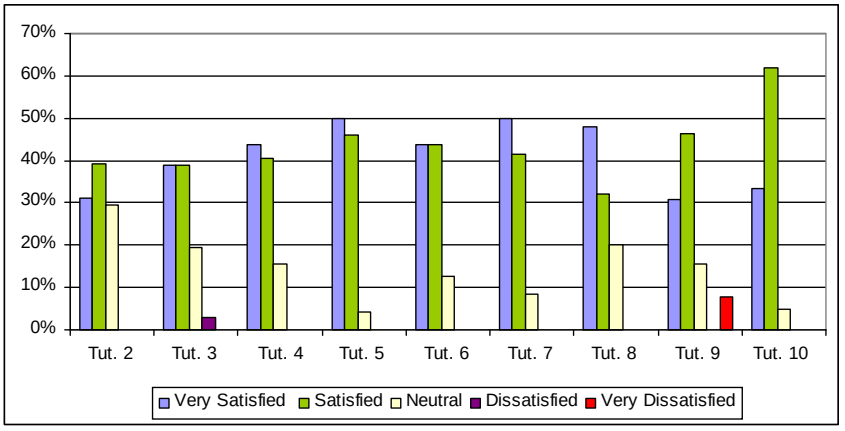


Figure 8.8 Survey on PP satisfaction

In terms of confidence level, students reported their confidence level when working with a partner by answering the question "*How do you rate your level of confidence solving the exercises with your partner?*", measured on a scale from 1 (very low) to 5 (very high). Figure 8.9 shows students' responses. Most students (on average 21 out of 28 students, 75%) responded that their level of confidence solving the tasks with their partner was high. Approximately 40% of the students in each tutorial session reported very high confidence with their ability to solve the exercises in pairs.

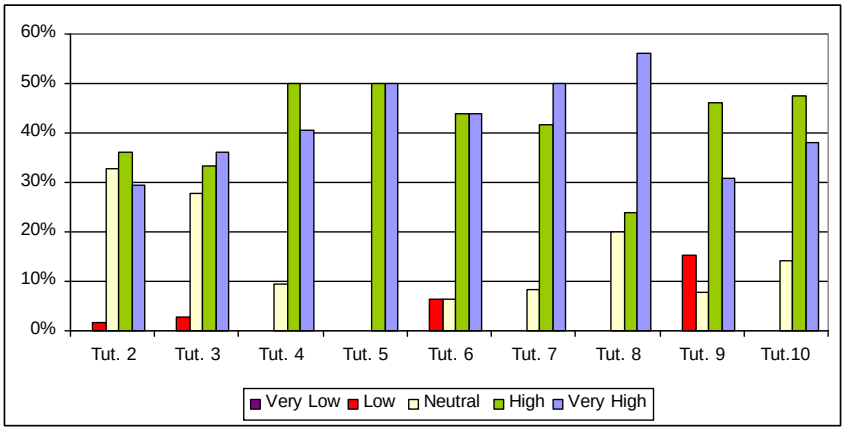


Figure 8.9 Survey on PP confidence

Data analysis on comparing the satisfaction and confidence levels between different groups of Conscientiousness was carried out using a non-parametric test known as Kruskal-Wallis Test. This statistical test was chosen because the dependent variables satisfaction and confidence were measured on an ordinal scale. Table 8.11 shows the mean satisfaction rank of paired students according to the Conscientiousness level. In this table, the group with the highest mean rank had the highest level of satisfaction. In five out of nine tutorials (tutorial 6 until tutorial 10), paired students of low Conscientiousness appeared to score higher satisfaction ranks; however the differences between the groups were mostly not significant. Only two tutorials (Tutorials 3 and 9) showed statistically significant differences between the three groups (p=0.01 and p=0.02 respectively) but overall results demonstrated that the

satisfaction levels of paired students were not affected by students' Conscientiousness levels. Similarly, we observed no significant differences in terms of confidence levels among the different Conscientiousness levels (see Table 8.12).

Table 8.11 Mean rank for satisfaction level

	Consc. Level	N	Mean Rank	Sig.
Tut. 2 N=61	Low	17	26.65	0.20
	Medium	23	35.76	
	High	21	29.31	
Tut. 3 N=36	Low	13	12.19	0.01
	Medium	11	20.55	
	High	12	23.46	
Tut. 4 N=23	Low	12	11.88	0.64
	Medium	4	9.75	
	High	7	13.50	
Tut. 5 N=22	Low	4	12.0	0.84
	Medium	11	12.0	
	High	7	10.43	
Tut. 6 N=13	Low	4	9.25	0.32
	Medium	5	6.30	
	High	4	5.63	
Tut. 7 N=12	Low	2	9.50	0.33
	Medium	6	6.40	
	High	5	5.40	
Tut. 8 N=21	Low	3	16.00	0.22
	Medium	10	9.60	
	High	8	10.88	
Tut. 9 N=10	Low	4	8.50	0.02
	Medium	3	4.50	
	High	3	2.50	
Tut.10 N=19	Low	4	11.5	0.32
	Medium	9	11.0	
	High	6	7.50	

Apart from measuring paired students' satisfaction and confidence, students' feedback on their overall pairing experience was also gathered. Thus data for the following questions were gathered from each tutorial:

"I felt that working with this partner was a productive experience." (Q1)

"I enjoyed working with my partner." (Q2)

"My motivation level increased when working with my partner." (Q3)

Table 8.12 Mean rank for confidence level

	Consc. Level	N	Mean Rank	Sig.
Tut. 2 N=61	Low	17	27.18	0.18
	Medium	23	36.04	
	High	21	28.57	
Tut. 3 N=36	Low	13	13.15	0.04
	Medium	11	19.86	
	High	12	23.04	
Tut. 4 N=23	Low	12	12.00	0.98
	Medium	4	11.50	
	High	7	12.29	
Tut. 5 N=22	Low	4	11.75	0.74
	Medium	11	12.27	
	High	7	10.14	
Tut. 6	Low	4	9.63	0.11

N=13	Medium High	5 4	7.00 4.38	
Tut. 7 N=12	Low Medium High	2 6 5	9.50 6.40 5.40	0.33
Tut. 8 N=21	Low Medium High	3 10 8	15.50 9.20 11.56	0.21
Tut. 9 N=10	Low Medium High	4 3 3	8.50 4.50 2.50	0.02
Tut.10 N=19	Low Medium High	4 9 6	11.50 11.33 7.00	0.23

Our analysis showed that on average 24 out of 28 students (86%) responded that their experience working with a partner was productive (see Figure 8.10). In the written feedback, one of the students mentioned the benefit of PP as "*helped each other out with different viewpoints*". Another student commented that PP helped him learnt a lot. In terms of enjoyment (Q2), students' responses are shown in Figure 8.11. Most students (on average 24 out of 28, 86%) agreed that working in pairs was an enjoyable experience. Similarly, students responded that their motivation level increased when working with a partner. On average 22 out of 28 students (78%) agreed with the statement Q3 (see Figure 8.12).

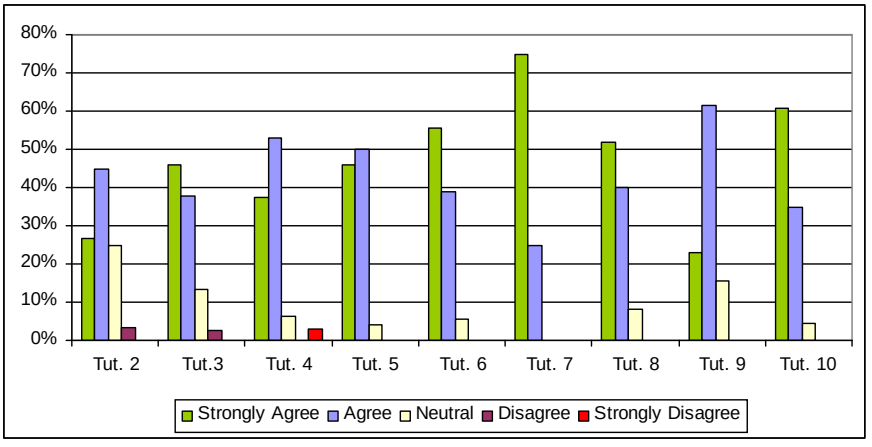


Figure 8.10 PP as a productive experience (Q1)

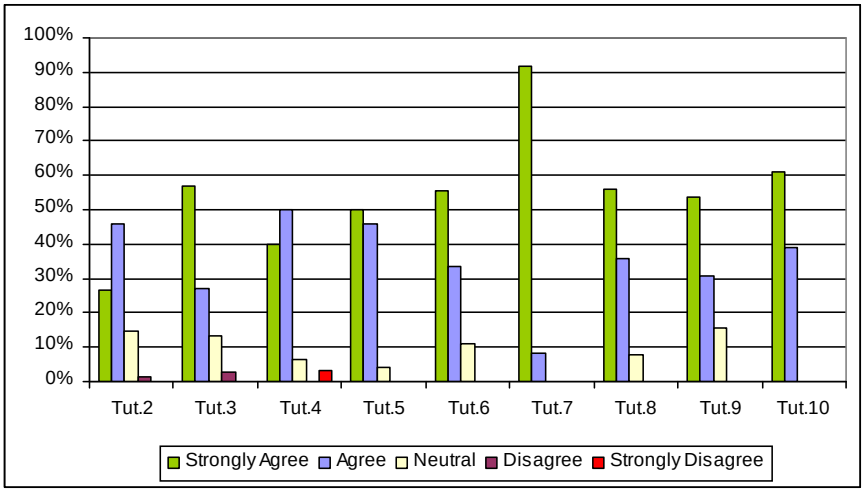


Figure 8.11 Enjoyment (Q2)

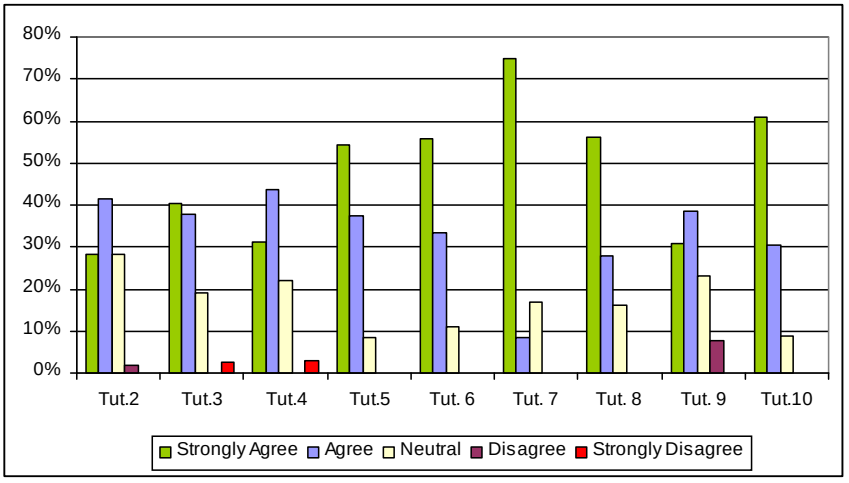


Figure 8.12 Motivation level (Q3)

8.7 Discussion

In this experiment, the effects of different Conscientiousness levels on PP's effectiveness was investigated using an advanced CS course attended by more mature students compared with our two previous experiments (see Chapters 6 and 7). The comparison of characteristics between the three experiments that looked at the effects of Conscientiousness on paired students' academic performance is presented on Table 8.13.

Similar to the findings from the 2nd experiment, we did not observe any significant correlation between the personality trait Conscientiousness and academic performance in the present experiment (*Exp 3*). Our hypothesis testing indicates that results significantly differed only for the midterm test scores, where paired students of high Conscientiousness levels achieved higher scores than the other groups. Nevertheless these differences were absent for the other dependent variables (i.e. assignments and final exam). Thus, results in the present experiment showed a lack of evidence to support our alternative hypothesis, except for on the

midterm test. This evidence suggests that the effects of Conscientiousness on paired students' academic performance may be trivial regardless of the nature of courses or subjects.

Table 8.13 Comparison of the three formal experiments investigating Conscientiousness trait

Experiment:	Exp 1 (Chapter 6)	Exp 2 (Chapter 7)	Exp 3 (Chapter 8)
Semester:	Summer 2009	Semester 1, 2009	Semester 1, 2009
Sample Size (N):	48	212	77
Course:	CS101	CS101	CS230
Subjects:	First Year undergraduate	First year undergraduate	2nd year undergraduate CS students
Tutorial settings:	Compulsory (2 hours in closed-lab)	Compulsory (2 hours in closed-lab)	Optional (An hour in closed-lab)
Alternative Hypothesis supported? (Yes/No)	No	No	No (except for the midterm test)
Correlation with Academic performance (Yes/No)	No (except for assignments)	No	No
Effect Size:	0.08*	0.07 (assignments) 0.06 (midterm) 0.04 (final exam)	0.11 (assignments) 0.32 (midterm) 0.10 (final exam)
Statistical Power:	0.28	0.14 (assignments) 0.10 (midterm) 0.08 (final exam)	0.12 (assignments) 0.71 (midterm) 0.11 (final exam)

** Effect size was generated based on the linear combination of DVs using MANOVA*

As a result of the non-statistical significance of our results in *Exp 3*, we carried out a statistical power analysis based on classes of parameters employed in this experiment (i.e. the sample size, alpha, and the observed effect size). We found that there was a very low statistical power obtained for the effects of treatment on two dependent variables (assignments and final exam). The overall low statistical power observed in this experiment is similar to the power generated from our 2nd experiment (see Table 8.13). Increasing a sample size in future study would have helped statistical power of the results (Dyba et al., 2006).

In the present experiment the statistical power for the midterm ANOVA test was found to be high due to the medium effect size observed for this particular test. Based upon the midterm ANOVA test, given the sample size employed (77), and a medium-sized estimate of the effect size (0.32) at α=0.05, a reasonably high statistical power was obtained (0.71). According to one of the course lecturers, the questions set for the midterm test were more difficult than the final exam questions. In this scenario, there is a possibility that the level of difficulty or complexity of the test may have influence the results, such that more conscientious students (i.e. eventually hard working students) tended to perform better on the more difficult test questions.

Our correlation analysis showed a weak correlation between the Conscientiousness trait and academic performance. Of the five traits, Openness to experience was the only trait that showed a significantly positive correlation with performance in both midterm test and the final exam. These results were consistent with our previous findings in the two experiments reported in Chapters 6 and 7, and also corroborate results reported in the educational-psychology literature (Farsides & Woodfield, 2003; Chamorro-Premuzic & Furnham, 2008;

Fruyt & Mervielde, 1996). According to personality psychologists, the Openness to experience is the trait that relates to attributes such as being intelligent, imaginative, and perceptive (McCrae & John, 1992; Digman, 1990). Thus one of our experiments (see Chapter 10) focused on whether this trait had a significant influence on the academic performance of students who paired programmed.

Based on the students' feedback on the surveys, we found that students were highly satisfied when working in pairs and their confidence level in solving the tasks were also increased. These results corroborate the previous findings in the experiments reported in Chapter 6 and 7.

8.8 Threats to the Validity

One major threat to the validity of our results is related to the issue of sampling bias. In this experiment, we had to rely on personality data from only the students who filled out the online IPIP-NEO test. Thus the sample could not be considered as completely random, but rather more like a self-selected sample. This situation can bias the results obtained due to some undesirable tendencies. For instance, those students who chose or were willing to give responses on their personality profile may be more obedient or more motivated to take part in the study and they may not be representative of the population.

In this experiment students' academic performance was used as a surrogate measure of PP's effectiveness. The use of surrogate measurement data for assessing the PP's effectiveness is also a potential source of threat. This is because students' academic performance may also be affected by their cognitive ability or learning strategy; thus they may perform well regardless of their attendance to the tutorial session. It is important to note however, that students attending the tutorials were typically those who need help from the tutor in order to get better understanding of the subject matter. Thus we believe that the pairing session during the tutorial somehow helped the students in improving their learning which eventually reflected in their performance in the test and exam.

Another limitation is that the tutorials were run by two different tutors assigned to this course; each of them was responsible for a block of tutorials of four weeks. This may have influenced the students' motivation to attend the tutorial in case a tutor's ability did not meet their expectation. Nonetheless, since the level of attendance varied weekly regardless of who did the tutoring, and that the background of tutors appointed for this course was similar, we believe that this issue may not significantly affect the results obtained in this experiment. Also, the course was taught by two different instructors throughout the semester, each handling a four-week block of teaching. Thus, differences in teaching style or delivery method may have had an influence on students' motivation and level of comprehension in this course.

8.9 Summary

In this experiment, we did not find support for our alternative hypothesis on the effects of Conscientiousness on paired students' academic performance on an advanced level CS

course. Although one of our statistical tests showed significance findings (i.e. midterm test scores were affected by the Conscientiousness levels), the same effects were absent in other performance measures (i.e. assignments and final exam). Therefore, the data did not provide results that were enough to refute (reject) the null hypothesis, except for the midterm test.

Due to the overall low level of statistical power observed for this experiment, with a low estimate of effect size, the experiment had only a little power to detect the treatment variance – a result that was consistent with our previous experiments. A general recommendation would be to possibly conduct an experiment with a larger sample size as it provides more stable estimates of the population parameter and increase the chances to detect effects.

In general, students gave positive feedback about PP. We found that most students reported a high level of satisfaction (79%) and confidence (75%) when working with their partners. These results were also consistent with our previous experiments. We also observed a significant positive correlation between the Openness to experience trait and academic performance measures. Thus, one of our future experiments will be investigating the effects of this trait on PP's effectiveness.

Chapter 9

THE FOURTH EXPERIMENT

This chapter describes an experiment conducted at the University of Auckland during the second semester of 2009. The subjects who participated in the experiment were first year undergraduate students enrolled in an introductory programming course. The experiment's objectives and details are explained in the following sections. Finally, the results obtained are discussed and the threats to the validity of our findings are also identified.

9.1 Experimental Objectives

The experiment described herein aimed to investigate the effect of the personality trait of Neuroticism on the academic performance of students practicing PP throughout one academic semester. Neuroticism is one of the FFM's personality factors reported to have a prominent role in learning and on educational contexts (De Raad & Schouwenburg, 1996). Neuroticism relates to the level of emotional stability, where high Neuroticism reflects a person's negative disposition such as feeling anxiety, hostility, or self-consciousness (Costa & McCrae, 1995). In contrast, a person who is low in Neuroticism exhibits a more resilient character represented by being composed, calm, and rarely discouraged (Costa & McCrae, 1995; Ogot & Okudan, 2006).

In a review of personality in learning and education by De Raad and Schouwenberg (1996), they mentioned that "*particularly at the University level, highly neurotic students are probably handicapped as compared with low neurotics.*" (p. 326). Thus we believe that this trait may play a role in determining the performance of students who pair programmed. Therefore, the aim of this experiment is to investigate whether or not Neuroticism plays a role in differentiating the performance of students who pair programmed. The goal definition for the formal experiment is the following (Basili, Shull, & Lanubile, 1999):

Object of study: PP technique.

Purpose: To improve the effectiveness of PP as a pedagogical tool in higher education institutions.

Focus: To investigate the influence of Neuroticism factor of FFM personality model that can potentially affect the success of the PP practice in a CS/SE course.

Perspective: From the point of view of the researcher.

Context: In the context of undergraduate CS/SE students.

9.2 Experimental Context

The formal experiment was conducted during the second semester of 2009, and participants were first year undergraduate students attending an introductory course for learning an

object-oriented programming language in Java, the Principles of Programming course (COMPSCI 101). During this course, students learnt about basic programming concepts and created a few small applications as part of their assignments. The teaching of this course consists of ten weeks of lectures and a compulsory weekly tutorial. The course instructor taught the basic concepts of programming during the one-hour lecture conducted three times per week.

The tutorials were held in a computer lab run by a tutor. During the tutorial, students who participated in the experiment worked in pairs when solving the exercises assigned by the tutor. Each of the tutorial sessions was treated as an independent experiment thus students' feedback regarding the pairing experience was gathered on every session.

9.3 Hypothesis

Of the five personality constructs, Neuroticism (or lack of emotional stability) is the factor that deals with feelings of anxiety, self-consciousness, impulsiveness, and vulnerability (De Raad & Schouwenburg, 1996; Costa & McCrae, 1995). In two longitudinal studies of two British university samples, findings showed that Neuroticism was negatively and significantly related to academic performance, particularly for examination marks (Chamorro-Premuzic & Furnham, 2003a). Similar findings were reported in their replication study where the negative relationship between academic success and Neuroticism was observed as a result of anxiety and impulsiveness traits (Chamorro-Premuzic & Furnham, 2003b).

It should however be noted that there is some evidence from organizational psychology that in certain conditions anxiety and Neuroticism may actually facilitate performance (Burch & Anderson, 2008). On a positive side, emotional stability is consistently related to self-efficacy, which in turn, affects performance (Schmitt, 2008; Barrick et al., 1998). Barick et al. (1998) reports that teams comprised of more emotionally stable members (i.e. low Neuroticism) are likely to achieve higher performance when compared with teams that include even a single member who is emotionally unstable. Therefore, we posited that the level of Neuroticism may influence the academic performance of students practicing PP. Therefore, we have investigated the following hypothesis in our experiment:

H_O: Differences in levels of Neuroticism do not affect the academic performance of students who pair programmed.

which is contrasted by the following alternative hypothesis:

H_A: Differences in levels of Neuroticism affect the academic performance of students who pair programmed.

Table 9.1 shows the categorization of pairs according to students' level of Neuroticism. A pair (N_{High}, N_{High}) denotes a pair combination where both students have high levels of Neuroticism.

In addition to investigating the abovementioned hypothesis, this experiment also looked into the relationship between a student's personality score with their academic performance, and their level of satisfaction and confidence when working in pairs.

Table 9.1 Pair configuration

Neuroticism Level	Pairing Groups
High	Pair (N_{High}, N_{High})
Medium	Pair (N_{Med}, N_{Med})
Low	Pair (N_{Low}, N_{Low})

9.4 Variables

In this research, PP's effectiveness was measured using students' academic performance in assignments (15%), a midterm test (15%) and final exam (60%). Hence, academic performance was our dependant variable, and level of Neuroticism (low, medium, high) our independent variable. The levels of satisfaction and confidence were measured using a questionnaire where all questions employed a five-point Likert-scale. We employed the same set of instruments as in our previous experiment (see Chapter 7).

9.5 Experimental Procedure

The experiment took place during the weekly compulsory tutorial sessions, run by a tutor and a few teaching assistants. We followed the same procedure carried out in our previous experiments (see Chapter 7). Students' personality and demographic data were gathered at the start of the semester. An online version of the IPIP-NEO was used to measure students' personality against the FFM. The results of the personality profiling were then used to allocate partners. For this purpose, the scores on the Neuroticism trait were used to assign paired students in three possible groups, representing three different levels of Neuroticism: low, medium and high. The grouping of participants per Neuroticism level was done based on the distribution of scores for the Neuroticism traits (i.e. low – lowest 40%; medium – middle 30%, high – highest 30%). This was done in order to provide a more balanced number of subjects within each group.

In every tutorial, pairs were allocated randomly within each group, and in addition a "single-factor between-group design" (Morgan et al., 2004) was the research design employed. Every tutorial lasted for two hours, where the first 45 minutes were used by the tutor to explain the topic, and the remaining 75 minutes were allocated for students to solve the exercises in pairs.

At the end of each tutorial, students filled out a short questionnaire providing feedback about working with their partners. They indicated their satisfaction level working with their partner by answering the question "*Please rate how satisfied are you working with your partner*" on a scale from 1 (very dissatisfied) to 5 (very satisfied). Students reported their confidence level by answering the question "*How do you rate your level of confidence solving the exercises with your partner?*" on a scale from 1 (very low) to 5 (very high). The exercises given during the tutorials were graded, thus contributing towards their final grade. In addition, assignments and the midterm test were also graded, however completed individually.

The learning outcome the experiment measured was determined by the scores on three assignments, a midterm test and final exam. During the tutorial sessions, students were required to solve a minimum of two programming problems with the assigned partner. Since tutorials varied from week to week, the experiments were designed in such a way as to minimize the effect of confounding factors due to differences in the tasks and level of complexity of the exercises assigned to the students. Therefore, the tasks and exercises remained the same throughout each week.

9.6 Results and Analysis

This section describes the results of the experiment including the sample's demographics data. The interpretation of the results is presented under the discussion section and finally the potential threats to the validity of the results are also discussed.

9.6.1 Demographics

A total of 270 students were enrolled in the course. Of these, 202 were male students (74.8%), 68 females (25.2%); subjects' age ranged from 19 to 47 years old (the mode age = 19 years). Of the 81 students who answered the demographics questionnaire, 64 students (79%) did not have any previous work experience; however 14 students (17%) indicated their programming competency as above average. Only 77 out of 270 students (29%) declared Computer Science as their major. Of the 270 students, 118 students (44%) completed the personality test. Therefore, the sample size used in our analysis was of 118 students.

Based on the demographics data, subjects came from various ethnic backgrounds, the majority being Chinese (26 students, 32%) and NZ/Pakeha (26 students, 32%), followed by South Korean (7 students, 8.6%), Indian (6 students, 7.4%), Asian (5 students, 6%), and Sri Lankan (4 students, 5%).

9.6.2 Data Distribution

The boxplot in Figure 9.1 shows the distribution of personality data for all the five traits. The box represents the middle 50% of the scores, with the upper and the lower tails indicating the 75^{th} and 25^{th} percentiles, respectively. The distribution of personality scores was fairly similar between Extraversion, Agreeableness, and Conscientiousness. However, a strong positively skewed distribution is observed in the Openness to experience trait. In terms of the median scores, the highest median belongs to the Neuroticism, followed by Agreeableness, whereas the lowest median score belongs to Openness to experience. The outliers represent cases where students obtained very high scores for the Openness to experience trait (i.e. above 75).

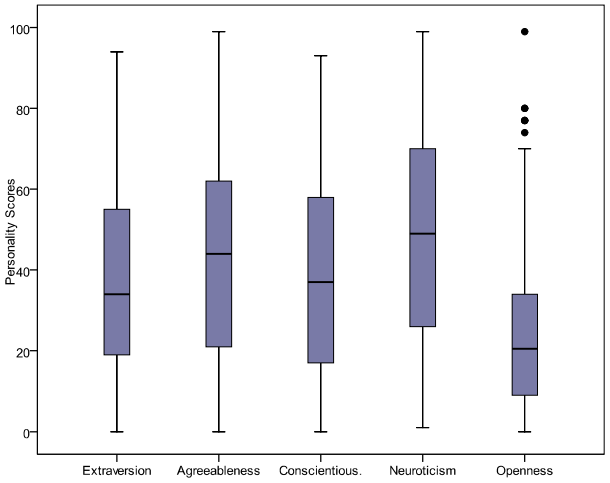


Figure 9.1 Comparison of FFM scores

Figure 9.2 shows the boxplot of assignment scores by Neuroticism levels. The boxplot showed that the spread or variation of scores is greatest in the medium Neuroticism group, followed by the low Neuroticism group. Data distributions for both the low and high Neuroticism groups were negatively skewed. In terms of the median scores, the highest median belongs to the high Neuroticism group, followed by the low Neuroticism group. The maximum possible score for the assignments was 15.

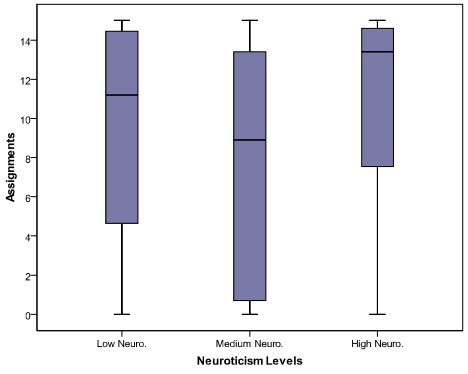


Figure 9.2 Comparison of assignments scores between groups

In terms of the midterm test scores, the flattest distribution was observed for the medium Neuroticism group, suggesting this group was more heterogeneous compared to the other two groups (see Figure 9.3). The median for the high Neuroticism group was slightly higher than the other two, whereas the lowest median value belongs to the medium Neuroticism group. The outliers in this boxplot showed cases where the midterm score was lower than 20%.

A similar pattern of distribution was observed regarding the final exam scores (see Figure 9.4). However, in this case, there was similar dispersion across the three groups, and there were no outliers. Of the three groups, data distributions were negatively skewed for the high

Neuroticism group and the highest median value belonged to this group. The maximum possible scores for midterm and final exam were both 100.

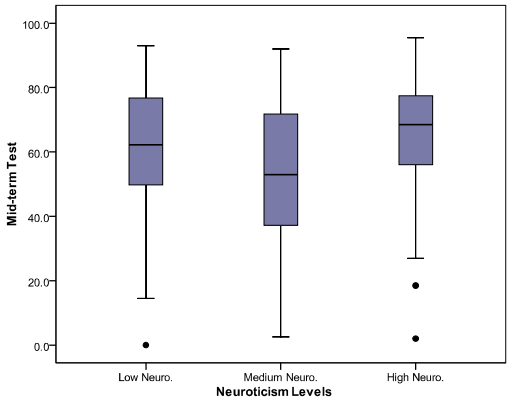


Figure 9.3 Comparison of midterm test scores between groups

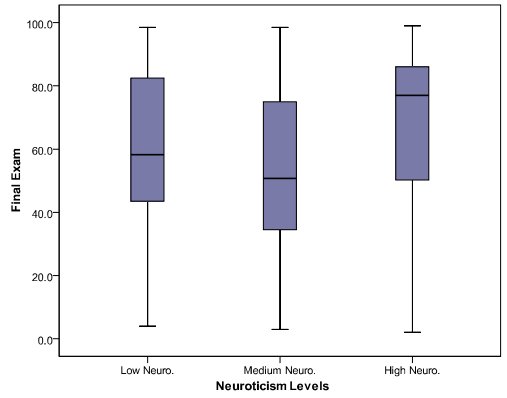


Figure 9.4 Comparison of final exam scores between groups

9.6.3 Correlational Analysis

In order to measure the strength of association between Neuroticism levels and academic scores we employed the Pearson's correlation coefficient (α = 0.05) (see Table 9.2). A parametric test was chosen because the sample size used in our experiment was not considered small (Pallant, 2007). No statistically significant association was found between Neuroticism and any measure of academic performance. As Table 9.2 shows, we also measured the level of association between the other four personality traits and students' academic performance. These results showed a statistically significant correlation between participants' Conscientiousness level and performance in assignments, and test scores ($r(116) = 0.19$, $p < 0.05$ for assignments, and $r(116) = 0.19$, $p < 0.05$ for the midterm test).

The positive significant correlation between Conscientiousness and assignment scores was consistent with those from our previous experiment (see Chapter 6). The findings regarding Conscientiousness also corroborated those reported in the psychology literature (e.g. Chamorro-Premuzic & Furnham, 2003b; Busato et al., 2000; Furnham et al., 2003) which suggest Conscientiousness as a consistent predictor for academic performance due to

the characteristics of highly conscientious individuals (i.e. persisting, achieving, grades orientation etc.)

Other than Conscientiousness, there is also a significant positive correlation between the midterm test score and Openness to experience ($r(116) = 0.23$, $p < 0.05$). This result also corroborates the results from our previous experiments (see Chapters 6 and 8).

Table 9.2 Correlation between academic performance and personality factors (N=118)

	Assign	Test	Final	Extrav.	Agreeab.	Consc.	Neuro.
Assign	1						
Test	0.56**	1					
Final	0.68**	0.83**	1				
Extrav.	-0.01	0.07	-0.02	1			
Agreeab.	-0.02	0.12	-0.02	0.09	1		
Consc.	0.19*	0.19*	0.15	0.28**	0.21*	1	
Neuro.	0.05	-0.01	0.01	-0.24*	-0.15	-0.26**	1
Openn.	0.01	0.23*	0.15	0.07	0.24**	0.01	0.21*

**. Correlation is significant at the 0.01 level (1-tailed).
*. Correlation is significant at the 0.05 level (1-tailed).

9.6.4 Hypothesis Testing

The hypothesis investigated herein was tested using the one-way between-group analysis of variance (ANOVA) ($\alpha = 0.05$) in order to compare the effects of the three levels of Neuroticism on paired students' academic performance. Table 9.3 shows the values for the mean scores and standard deviations for each Neuroticism level.

Table 9.3 Mean and standard deviation of paired students of different level of Conscientiousness

Performance Measures	Neuroticism Level	N	Mean	SD
Assignments (Range: 0 to 15)	Low Neuro.	45	9.71	5.35
	Medium Neuro.	43	8.47	5.45
	High Neuro.	30	11.21	4.64
	Total	118	9.64	5.28
Test Scores (Range: 0 to 100)	Low Neuro	43	60.87	20.58
	Medium Neuro.	42	52.44	22.56
	High Neuro.	30	64.35	22.76
	Total	115	58.70	22.26
Final Exam (Range: 0 to 100)	Low Neuro	42	59.62	23.86
	Medium Neuro.	40	52.00	27.10
	High Neuro.	29	64.10	30.99
	Total	111	58.04	27.22

Levene's test for homogeneity of variances assesses whether the population variances for each of the Neuroticism groups are significantly different form each other (Carver & Nash, 2006). The results from the Levene test (see Table 9.4) showed that the assumption of homogeneity of data was not violated ($F = 0.73$, $p = 0.48$ for assignments; $F = 0.24$, $p = 0.79$ for midterm test, and $F = 1.25$, $p = 0.29$ for final exam).

The ANOVA results (see Table 9.5) showed that at the $p < 0.05$ level there was no statistically significant difference in academic performance between the three groups of

Neuroticism (i.e. $F(2,115) = 2.45$, $p = 0.09$, for assignments; $F(2,112) = 2.93$, $p = 0.06$, for midterm test; $F(2,108) = 1.80$, $p = 0.17$, for final exam). Since none of the F values were statistically significant, no post-hoc analysis was needed. Our results indicated that we could not find strong support to reject the null hypothesis. Therefore, based on our data, we found that paired students' academic performance was not significantly affected by differences in Neuroticism levels.

Table 9.4 Test of Homogeneity of variances

	Levene Statistic	df1	df2	Sig.
Assignments	0.73	2	115	0.48
Test Scores	0.24	2	112	0.79
Final Exam	1.25	2	108	0.29

Table 9.5 ANOVA results

		Sum of Squares	df	Mean Squares	F	Sig.
Assign.	Between Groups	133.39	2	66.69	2.45	0.09
	Within Groups	3135.04	115	27.26		
	Total	3268.43	117			
Test	Between Groups	2806.18	2	1403.09	2.93	0.06
	Within Groups	53679.72	112	479.28		
	Total	56485.90	114			
Final Exam	Between Groups	2630.18	2	1315.09	1.80	0.17
	Within Groups	78897.59	108	730.53		
	Total	81527.77	110			

9.6.5 Statistical Power Analysis

Due to the non-significance of our findings we conducted a post hoc power analysis to compute the statistical power of the ANOVA test employed in our experiment. Similar to our previous experiments, we used G*Power for the analysis (Version 3.1.2) (Faul et al., 2007). This analysis was carried out separately for each of the dependent variables using the F-test family of the one-way ANOVA. Tables 9.6 to 9.8 show the protocols of the power analysis where the input and output parameters were specified. Considering the standard convention of effect size defined by Cohen (1988), we found that there is a nearly medium range of effect size for dependent variables assignments and midterm test (i.e. 0.20, and 0.22 respectively, see Tables 9.6 and 9.7). With these effect sizes and the given sample size, the power generated was 0.47 and 0.54 respectively, which were below the recommended 0.80 (Cohen, 1988; Dyba et al., 2006).

A lower statistical power was also observed for the dependent variable final exam (i.e. 0.36, see Table 9.8). Statistical power represents the likelihood that a treatment effect will be observed whenever there is one. Given our analysis presented low statistical power relating to all the dependent variables, it is inappropriate to suggest that our results would correspond to what would be most likely to occur in a higher education environment when students are paired according to their Neuroticism level. In other words, the statistical power of our results is not sufficient to provide confidence that these results correspond to what would be most

likely observed when we investigate the relationship between academic performance and Neuroticism levels in paired students in higher education settings.

Table 9.6 Power analysis protocol (assignments)

F tests – ANOVA: Fixed effects, omnibus, one-way			
Analysis:	Post hoc: Compute achieved power		
Input:	Effect size f	=	0.20
	α err prob	=	0.05
	Total sample size	=	118
	Number of groups	=	3
Output:	Noncentrality parameter λ	=	4.73
	Critical F	=	3.07
	Numerator df	=	2
	Denominator df	=	115
	Power (1-β err prob)	=	0.47

Table 9.7 Power analysis protocol (midterm test)

F tests – ANOVA: Fixed effects, omnibus, one-way			
Analysis:	Post hoc: Compute achieved power		
Input:	Effect size f	=	0.22
	α err prob	=	0.05
	Total sample size	=	115
	Number of groups	=	3
Output:	Noncentrality parameter λ	=	5. 64
	Critical F	=	3.08
	Numerator df	=	2
	Denominator df	=	112
	Power (1-β err prob)	=	0.54

Table 9.8 Power analysis protocol (final exam)

F tests – ANOVA: Fixed effects, omnibus, one-way			
Analysis:	Post hoc: Compute achieved power		
Input:	Effect size f	=	0.18
	α err prob	=	0.05
	Total sample size	=	111
	Number of groups	=	3
Output:	Noncentrality parameter λ	=	3.55
	Critical F	=	3.08
	Numerator df	=	2
	Denominator df	=	108
	Power (1-β err prob)	=	0.36

Figure 9.5 shows the graphs of power as a function of the effect size *f* for the three different significance levels (α= 0.05; α= 0.1; α= 0.15). In order to achieve the desired high level of statistical power (i.e. 0.80) for the specified parameters (number of groups = 3; sample size = 118) the effect size should be of at least medium size (i.e. greater than 0.27)

for a significance level of 0.05. Similarly, a medium effect size is required to yield a high level of power (i.e. 80%) for an increased alpha level (i.e. $\alpha \geq 0.1$).

The plots in Figure 9.6 visualize the relationship between statistical power and sample size when the sample size changes from 50 to 500, at the effect size of 0.20. The statistical power for alpha level 0.15 is more sensitive to sample size than those for $\alpha = 0.05$ and $\alpha = 0.1$. In order to obtain the recommended statistical power of 0.80, it can be seen from the plot that this would have required a sample size a little over double than the current sample size (i.e. 250 compared to 118). Therefore, increasing sample size in future replication study may help increase the statistical power.

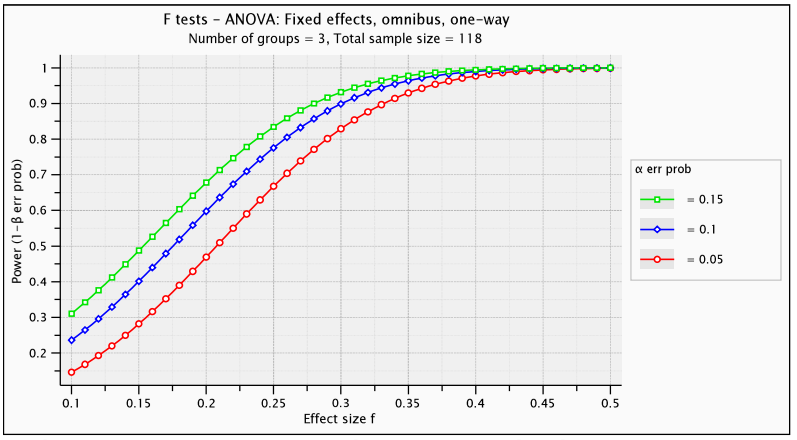


Figure 9.5 Power as a function of effect size (F Test – ANOVA)

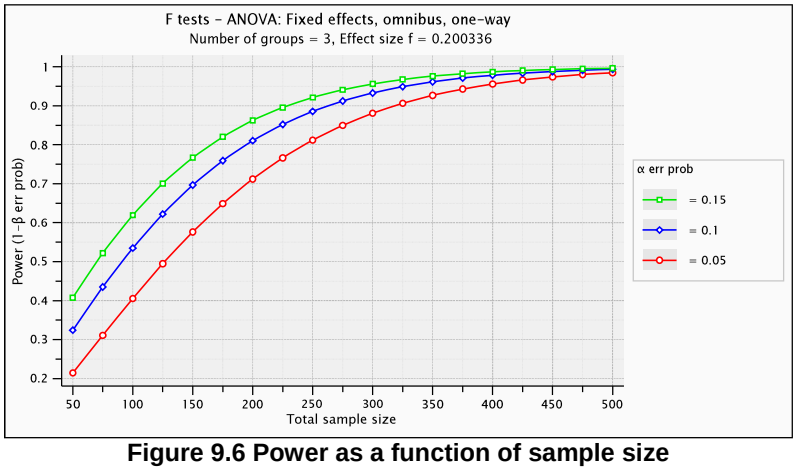


Figure 9.6 Power as a function of sample size

9.6.6 Results on Satisfaction and Confidence

We gathered data on paired students' satisfaction and confidence when working with their partner using a questionnaire distributed during the tutorial sessions. These data were gathered for eight weeks of tutorial classes. We did not gather the data for the first two weeks of classes in order to give students ample time to familiarize themselves with PP. Data were analyzed separately as each tutorial was treated as a single independent "mini-experiment". The dependent variable satisfaction was measured on a scale from 0 (*very dissatisfied*) to 5 (*very satisfied*); and confidence level was measured on a scale from 0 (*very low*) to 5 (*very*

high). The response rate of the post-experimental questionnaire was approximately 42% for every tutorial (excluding the Tutorial 3). On average, 60 (85.7%) out of an average of 70 students attending the tutorials, were satisfied working with their partner (see Figure 9.7).

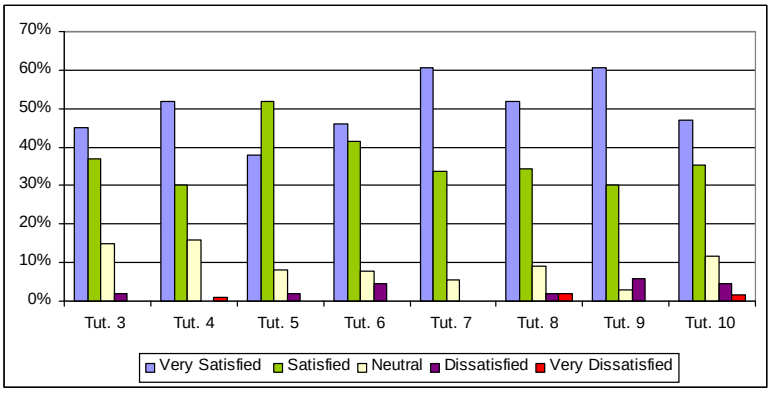


Figure 9.7 Survey on PP satisfaction

Table 9.9 shows the mean rank for the Kruskal-Wallis test (α = 0.05), which was used to compare the satisfaction rates between Neuroticism levels. The group with the higher rank indicates the higher satisfaction rates. Overall, results demonstrated that the satisfaction levels of paired students were not affected by different levels of Neuroticism. Of eight weeks of tutorials, only the last tutorial (tutorial 10) showed a significant difference in satisfaction across the three levels of Neuroticism: χ^2 (2, 68)=13.12, *p*=0.001. Nevertheless, these data also showed the trend that, according to their mean rank, paired students with low Neuroticism had higher satisfaction compared with the other two Neuroticism levels.

Table 9.9 Mean rank for satisfaction level

	Neuro. Level	N	Mean Rank	Sig.	Satisfied/ Very Satisfied (%)
Tut. 3* N=46	Low Medium High	16 20 10	25.00 25.30 17.50	0.22	82.6
Tut. 4 N=95	Low Medium High	29 35 31	51.60 48.67 43.87	0.48	83.2
Tut. 5 N=63	Low Medium High	18 28 17	34.61 29.11 34.00	0.45	90.5
Tut. 6 N=65	Low Medium High	23 24 18	37.22 32.94 27.69	0.22	87.7
Tut. 7 N=71	Low Medium High	25 23 23	32.84 38.00 37.43	0.54	94.4
Tut. 8 N=54	Low Medium High	15 20 19	31.10 26.80 25.39	0.49	87.0
Tut. 9 N=69	Low Medium High	25 27 17	37.80 34.94 30.97	0.46	91.3
Tut.10 N=68	Low Medium High	20 18 30	45.30 36.06 26.37	0.00	82.4

Note (*) Data were collected only for the last two days of the tutorial for that week

In terms of confidence levels, approximately an average of 59 (84.3%), out of an average of 70 students attending the tutorial reported high confidence (both high and very high) in solving the tasks with their partner (see Figure 9.8). Table 9.10 shows the mean rank for paired students' confidence level based on the analysis of the returned surveys. There was only one tutorial that presented a significant difference of confidence level across the three groups (tutorial 4): χ^2 (2, 95)=10.69, p=0.01. This particular result indicates that the low Neuroticism group obtained the highest confidence level compared with the other two groups. Overall, we found that confidence in solving the exercises was generally high among the low and medium Neuroticism groups. These results suggest a tendency of students of lower or moderate Neuroticism to believe in the correctness of their programming solutions compared to the high Neuroticism pairs.

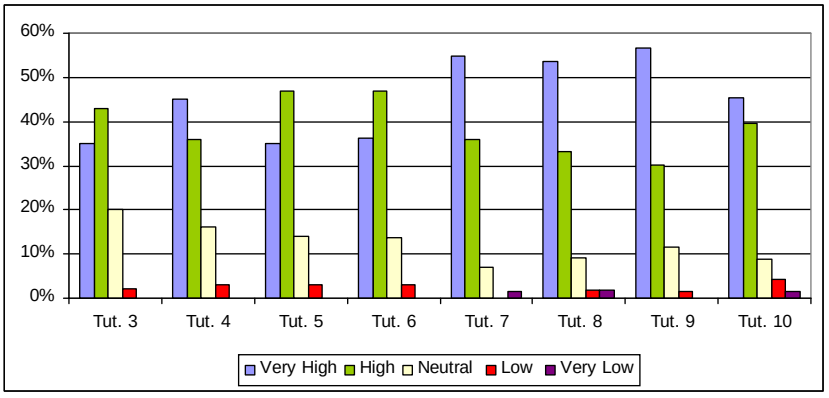


Figure 9.8 Survey on PP confidence

Table 9.10 Mean rank for confidence level

	Neuro. Level	N	Mean Rank	Sig.	% of high confidence
Tut. 3 N=46	Low Medium High	16 20 10	25.19 25.80 16.20	0.12	78.3
Tut. 4 N=95	Low Medium High	29 35 31	56.86 51.14 36.16	0.01	81.1
Tut. 5 N=63	Low Medium High	18 28 17	35.58 29.32 32.62	0.46	82.5
Tut. 6 N=65	Low Medium High	23 24 18	38.13 31.77 30.08	0.29	83.3
Tut. 7 N=71	Low Medium High	25 23 23	33.76 38.89 35.54	0.62	91.6
Tut. 8 N=54	Low Medium High	15 20 19	29.97 28.10 24.92	0.57	88.9
Tut. 9 N=69	Low Medium High	25 27 17	33.54 37.06 33.88	0.74	86.9
Tut.10 N=68	Low Medium High	20 18 30	41.20 35.06 29.70	0.09	85.3

In addition to measuring paired students' satisfaction and confidence levels, their responses to the following questions were also gathered:

"I felt that working with this partner was a productive experience." (Q1)

"I enjoyed working with my partner." (Q2)

"My motivation level increased when working with my partner." (Q3)

On average, 63 out of 70 students (90%) responded that their experience working with the partner was productive (see Figure 9.9). In terms of enjoyment (Q2), students' responses are shown in Figure 9.10. Most students (on average 62 out of 70, 88.5%) agreed that working in pairs was an enjoyable experience. Similarly, students responded that their motivation level increased when working with their partner. On average 59 out of 70 students (84.3%) agreed with statement in Q3 (see Figure 9.11).

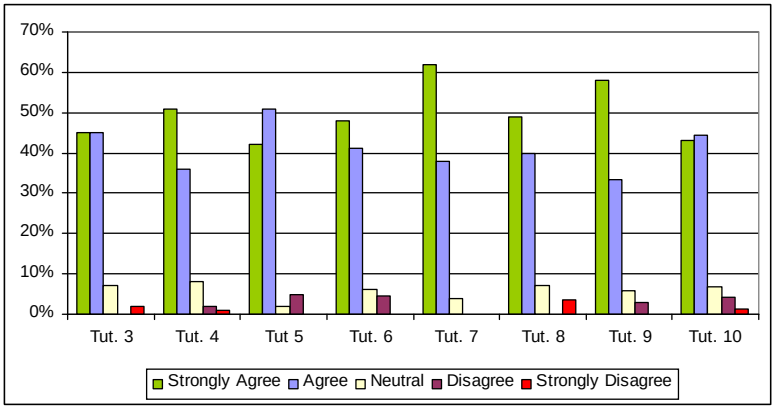


Figure 9.9 PP as a productive experience (Q1)

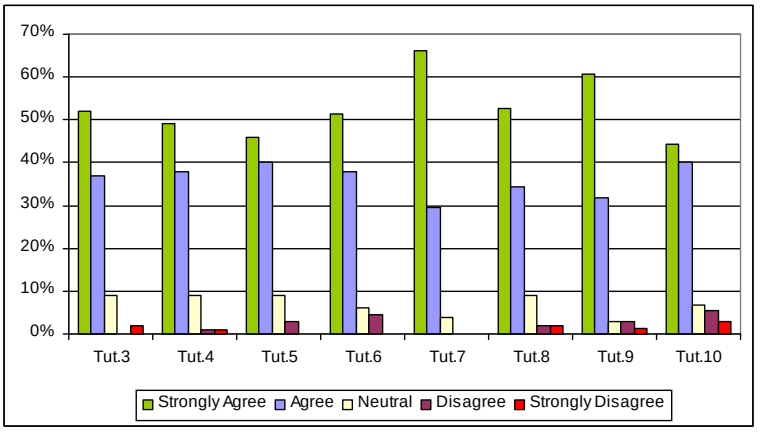


Figure 9.10 Enjoyment (Q2)

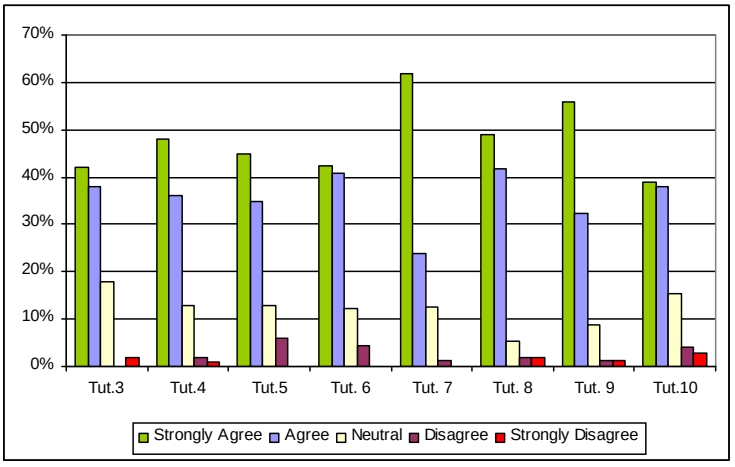


Figure 9.11 Motivation level (Q3)

9.7 Discussion

Although Neuroticism is reported in some studies to be related with the tendency to have "poorer" performance (e.g. Barrick et al., 1998; Chamorro-Premuzic & Furnham, 2003b; Bell, 2007) the findings from our experiment do not support this view. Based on the ANOVA analysis, we did not find any significant differences in academic performance between paired students of different Neuroticism levels. These results are consistent with other findings from previous research linking Neuroticism to academic performance among students in tertiary institutions (e.g. Furnham et al., 2003, Busato et al., 2000; Farsides & Woodfield, 2003).

In regards to the relationship between personality and team performance, a meta-analysis by Peeters et al. (2006) suggests that the elevation in emotional stability (i.e. low Neuroticism) is not significantly related to team performance due to the "broad concept" or wider impression represented by this trait. Instead, they proposed that the facets within the Neuroticism trait (e.g. self-consciousness, impulsiveness) should be empirically tested in order to obtain a more genuine effect (Peeters et al, 2006). Thus, future replication studies may consider investigating low-level facets of Neuroticism.

Existing research evidence also suggests that possible moderator effects could potentially influence the personality-team relationship (Bell, 2007; Peeters et al., 2006). One such effect is the type or the complexity of the task engaged in by the team (Bowers et al., 2000; Peeters et al., 2006). Bowers et al. (2000) suggest that the homogeneity of personalities of team members had very little effect on team performance, particularly on low-difficulty tasks. Thus, one of the possible reasons for the lack of statistically significant findings in this experiment could be related to the less complex tasks assigned to students. Future research should investigate mediator variables in order to better understand the impact of personality traits on performance. For example, a qualitative study on the nature of collaboration in PP by Walle & Hannay (2009) revealed some relationships between personality traits and the type of collaborations that may affect pair performance.

The significant positive correlation demonstrated in our findings between Conscientiousness and performance in assignments and midterm test indicates the tendency that students' Conscientiousness level may play a role in predicting performance. This is because Conscientiousness is the one consistently bearing significant positive relationship with high achievement of academic as well as team performance as also reported in the literature (Barrick et al., 1998; English et al., 2004; Peeters et al., 2006; Poropat, 2009). Nevertheless, our previous experiments do not seem to support this evidence (see Chapter 7). The significant positive correlation between the midterm test and Openness to experience was consistent with findings from our previous experiments (see Chapters 6 and 8).

In terms of the satisfaction level, overall results showed that differences in Neuroticism levels were not significant in affecting students' contentment while working in pairs. Despite these results, lower Neuroticism pairs scored higher satisfaction in most tutorials compared with the other groups (see Table 9.9). This perhaps relates to the common characteristic of low Neuroticism individuals being well adjusted people and likely to excel in team settings, as reported by Driskel et al. (2006).

9.8 Threats to the Validity

One of the potential threats to the internal validity of this experiment relates to the issue of changing partners during the tutorial. Some students failed to turn up to their allocated tutorial and attended a different session without informing the tutor. This could have led to students being paired with students from different Neuroticism groups. However, according to the tutor, these uncontrolled circumstances occurred sparingly thus minimizing the potential to bias the results.

Another potential threat relates to gender differences. As reported by Schmitt (2008), the interaction of Neuroticism and gender had significant impact on self-efficacy and performance. In this experiment, approximately 75% of the subjects are male students; therefore we believe that the probability of such a significant impact would be minimal due to the lower number of females enrolled in the course.

Another limitation refers to the fact that the performance measures used in this experiment may also be affected by levels of cognitive ability. In this experiment we used students' academic performance as surrogate measures of PP's effectiveness. Thus, there is a possibility that performance is affected by students' ability and competency in programming. However, since the experiment aimed to improve students' learning due to practicing PP throughout the entire semester, measuring their academic performance is in our view appropriate to our context. In addition, empirical evidence shows that the predictive power of one's cognitive ability in association with academic performance is relatively low compared to personality traits (Furnham et al., 2003). Therefore, students' cognitive ability may not have affected the results presented herein.

9.9 Summary

In this experiment, the findings showed that, based on the sample employed, paired students' performance was not significantly affected by the different levels of Neuroticism. The lack of support for the alternative hypothesis could be attributed to the low complexity of tasks assigned to students, and perhaps the existence of moderator variables mediating the relationship between personality traits and performance. Regardless of any possible threats to the validity of the results, the lack of statistical significance might have been due to the lower statistical power observed in this experiment.

The findings from this experiment also indicate that students' satisfaction and confidence levels did not differ depending on the levels of Neuroticism when pairing. We also observed that Conscientiousness and Openness to experience appeared to be significantly associated with performance. These results were consistent with our previous experiments. The next chapter describes the experiment investigating the effects of the Openness to experience trait on PP's effectiveness.

Chapter 10

THE FIFTH EXPERIMENT

This chapter describes an experiment conducted at the University of Auckland during the first semester of 2010. The subjects involved in the experiment were first year undergraduate students enrolled in an introductory programming course. The objectives and details of this experiment are explained in the following sections. Finally, the results obtained are discussed and the threats to the validity of our findings are also identified.

10.1 Experimental Objectives

The objective of this experiment was to improve the effectiveness of PP as a pedagogical tool for CS/SE education by investigating the effects of the personality trait Openness to experience on the academic performance of students practicing PP. Openness to experience is one of the FFM's traits that reflects personality characteristics such as being curious, imaginative, original, and broadminded (Costa & McCrae, 1992b). An individual with high Openness to experience tends to be more creative and willing to experiment with new ideas compared with an individual who is less Open to experience, who in turn prefers to use conventional methods, or is unwilling to accept changes (Driskell et al., 2006; LePine, 2003). In the context of students practicing PP, Openness to experience may play a role in differentiating students' academic performance. The detailed goal definition for this experiment is outlined as follows (Basili, Shull, & Lanubile, 1999):

> *Object of study:* PP technique.
> *Purpose:* To improve the effectiveness of PP as a pedagogical tool in higher education institutions.
> *Focus:* To investigate the influence of the Openness to experience trait of the FFM personality model given the assumption that it may potentially affect the success of the PP practice in CS/SE course.
> *Perspective:* From the point of view of the researcher.
> *Context:* In the context of undergraduate CS/SE students.

10.2 Experimental Context

Our fifth experiment was conducted in the tutorial labs of an introductory undergraduate course – Principles of Programming (COMPSCI 101), where participants were first year undergraduate students. The teaching component of this course consisted of ten weeks of lectures and nine weeks of compulsory tutorials. The main aim of this course was to provide students with the basic concepts of object-oriented programming development in Java.

Lectures were given three times a week, each lasting for an hour; in addition, there was a two-hour tutorial session once per week, run by a tutor and a few teaching assistants. During the tutorials, students worked with their allocated partners; data about the students' pairing experience was gathered from every tutorial session. Students willing to participate in the experiment were required to sign a consent form to fulfill the ethical requirements of the University of Auckland's Human Participant Ethics Committee.

10.3 Hypothesis

Openness to experience (also known as *Intellect*) is the fifth factor of the FFM that relates to an individual's intellectual curiosity, needs for variety, and aesthetic sensitivity according to the person's cognitive, affective and behavioral tendencies (Costa & McCrae, 1995). The people who are high on Openness to experience are described as being imaginative, intellectual, receptive to new ideas, and also broad-minded (LePine, 2003; McCrae & John, 1992).

Personality research on team settings reports that teams composed of highly opened to experience members are able to develop a more diverse methods or alternatives in problem-solving tasks (LePine, 2003). It has also been reported that Openness to experience emerges as a strong predictor of team performance because the team members who scored high on this trait are more adaptable and capable of handling changes that occur in a dynamic environment (Bell, 2007). In an academic setting, Openness to experience has been positively correlated with undergraduate academic success, in particular to students' final grades (Farsides & Woodfield, 2003; Dollinger & Orf, 1991). The findings from our previous studies also showed a significant positive correlation between Openness to experience and students' academic achievement in the midterm test and in the final exam (see Chapters 6 – 9). Given this line of reasoning, we conjecture that paired students' academic performance may be influenced by the level of Openness to experience. Hence, the following hypothesis was proposed:

> *H_O: Differences in the level of Openness to experience do not affect the academic performance of students who pair programmed.*

Which is contrasted by the following alternative hypothesis:

> *H_A: Differences in the level of Openness to experience affect the academic performance of students who pair programmed.*

Table 10.1 shows the categorization of pairs according to students' level of Openness to experience. A pair (O_{High}, O_{High}) denotes a pair combination where both students have high levels of Openness to experience. This experiment compared the performance of students in these groups based on their academic achievement in the course. Our experiment also looked into the association between each student's personality score with their academic performance, level of satisfaction and confidence when working in pairs.

Table 10.1 Pair Configuration

Openness to experience level	Pairing groups
High	Pair (O_{High}, O_{High})
Medium	Pair (O_{Med}, O_{Med})
Low	Pair (O_{Low}, O_{Low})

10.4 Variables

In this experiment, PP's effectiveness was measured using students' academic performance in assignments (14%), a midterm test (15%), and final exam (60%). Hence, PP's effectiveness, satisfaction, and confidence were our dependent variables and level of Openness to experience (low, medium, and high) our independent variable. Level of satisfaction and confidence were measured using a questionnaire where all questions employed a five-point Likert-scale (see Appendix B.4). We used the same set of instruments as in our previous experiment (see Chapter 9).

10.5 Experimental Procedure

The experiment took place during the tutorial sessions. We followed the same procedure carried out in our previous experiments (see Chapters 6 – 9), where each of the tutorial sessions was treated as an independent formal experiment. Students' personality and demographic data were gathered during the first week of the semester. An online version of the IPIP-NEO inventory was used to measure students' personality against the FFM. The results of the personality profiling were then used to allocate partners. For this purpose, the scores on the Openness to experience trait (i.e. between 0 and 99) were used to assign paired students into three possible groups, representing the three different levels of Openness: low, medium and high. The grouping of participants per Openness level was done based on the distribution of scores for the Openness to experience traits (i.e. low – lowest 15%; medium – middle 20%, high – highest 65%). This was done in order to provide a more balanced number of subjects within each group.

In every tutorial, pairs were allocated randomly within each group. Thus a "*single-factor between-group design*" was the research design employed in this experiment (Morgan et al., 2004). Every tutorial lasted for two hours where the first 45 minutes were used by the tutor to explain the topic, and the remaining 75 minutes were allocated for students to solve the programming exercises in pairs. During the tutorial sessions, students were required to solve a minimum two programming problems with the assigned partner.

Before the end of every tutorial, students provided feedback relating to working with the partner by filling out a short questionnaire (see Appendix B.4). The exercises given during the tutorials were graded, thus contributing towards students' final grade. In addition, assignments and the midterm test were also graded, however completed individually.

The outcomes measured from the experiment were the students' academic performance in their three assignments, in a midterm test and in a final exam. Since tutorial exercises

varied from week to week, the experiments were designed in such a way to minimize the confounding factor which might occur due to differences in tasks and exercises' levels of complexity (Arisholm et al., 2007). Therefore, the same set of exercises was given throughout a week.

10.6 Results and Analysis

This section describes the results of the experiment including the subjects' demographics data. The results are interpreted in the discussion section and finally the potential threats to the validity of the results are also discussed.

10.6.1 Demographics

A total of 488 students were enrolled in the COMPSCI 101 course during the first semester of 2010. Of these, 372 (76.2%) were male students, and 116 (23.8%) were female students. The subjects' age ranged from 18 to 55 years old (the mode age = 19 years). Of the 164 students who answered the demographic survey, 138 (84.1%) did not have any work experience; however 55 (33.5%) students indicated that their programming competency was above average. Subjects came from various ethnic backgrounds: 59 (40%) NZ/Pakeha, 44 (26.8%) Chinese, 18 (11%) Indian, 10 Korean (6%) and other minority ethnic groups were Asian, European, and Pacific Islanders. Of the 488 students, 154 (31.6%) students completed the personality test and have consented to participate in the study. Of these 154 students, only 137 students remained enrolled throughout the semester and sat the midterm test and the final exam. Therefore, the sample size used in our analysis was 137 students.

10.6.2 Data Distribution

The boxplots in Figure 10.1 show the distribution of personality scores based on the FFM traits. As can be seen in Figure 10.1, the distribution of scores for Openness to experience is positively skewed compared with the distribution of other personality factors. The distributions of scores between Agreeableness and Conscientiousness have a similar spread, representing almost the same median value. The dispersion of scores between Extraversion and Neuroticism is also similar. The lowest median value belongs to the Openness to experience. The additional black dot in the distribution of Openness to experience is an outlier, representing a student who obtained a very high score in this personality trait.

Figure 10.2 shows the distribution of assignment scores according to students' level of Openness to experience. Each boxplot represents a negatively skewed distribution where the distribution of assignments' scores for the high Openness group was more peaked than that for both low and medium groups. Such a peaked distribution, however, did not seem to affect the results of most statistical analyses as mentioned by Morgan et al. (2004). Both low and medium Openness to experience groups showed a similar spread. The highest and lowest

medians were shown for the high Openness and medium Openness group, respectively. The outliers indicate cases where the students did not complete some of their assignments.

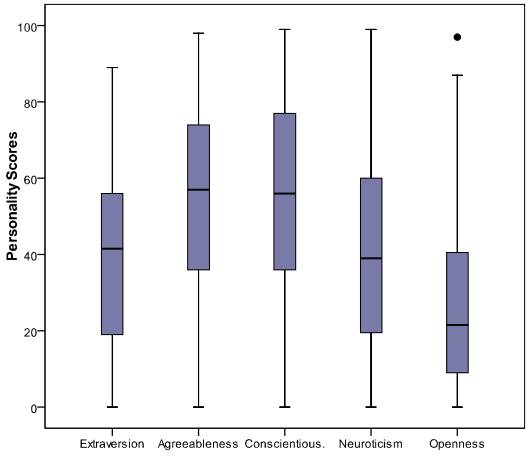


Figure 10.1 Comparison of FFM scores

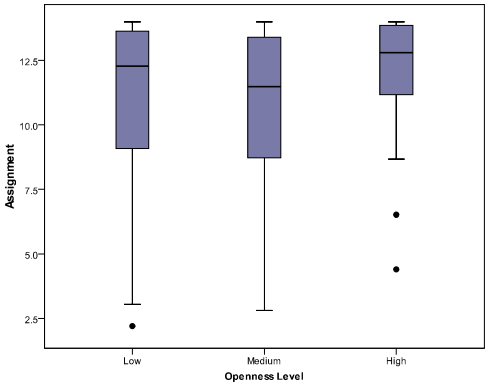


Figure 10.2 Comparison of assignments scores between groups

The boxplots in Figure 10.3 show the distribution of the midterm test scores for each of the Openness to experience levels. The dispersion of scores and median for both low and medium Openness to experience groups were similar and differed from the high Openness group, which showed a more peaked distribution, and the highest median overall.

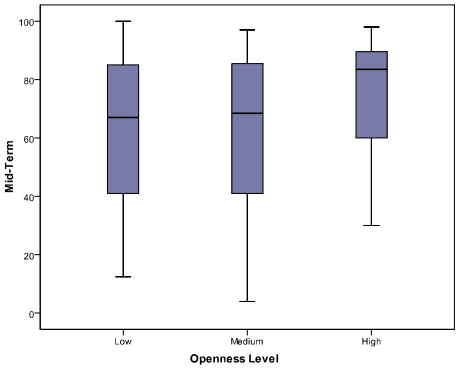


Figure 10.3 Comparison of midterm scores between groups

Figure 10.4 shows the distribution of scores for the final exam according to students' level of Openness to experience. The dispersion of scores and median for both low and medium Openness to experience groups were similar, and differed from the high Openness group, which showed a more peaked distribution, and the highest median overall.

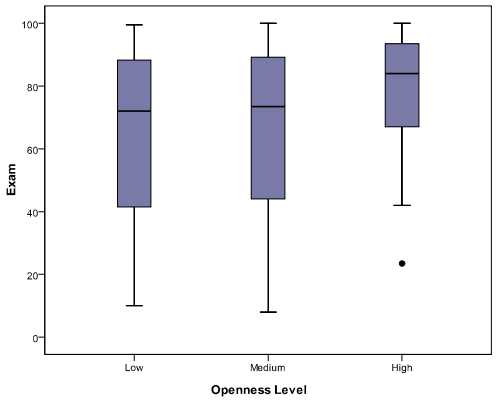


Figure 10.4 Comparison of final exam scores between groups

10.6.3 Correlation Analysis

A correlation analysis using the Pearson's correlation coefficient (α = 0.05) was performed to measure the strength of association between levels of Openness to experience and paired students' academic performance (see Table 10.2). The results showed a statistically significant positive correlation between Openness to experience and the midterm test ($r(137) = 0.18$, $p < 0.05$) and between Openness to experience and the final exam ($r(135) = 0.17$, $p < 0.05$). These findings corroborate the results from our previous experiments (see Chapters 6, 7, and 8). In addition, there was also a significant positive correlation between Conscientiousness and all performance measures (r (137) = 0.17, $p < 0.05$ for assignments; $r(137) = 0.19$, $p < 0.05$ for the midterm test; r(135) = 0.18 for the final exam). These findings were partly consistent with those from our previous experiment (see Chapter 9).

Table 10.2 Correlation between academic performance and the FFM (N=137)

	Assign	Test	Final	Extrav.	Agreeab.	Consc.	Neuro.
Assign	1						
Test	0.68**	1					
Final	0.69**	0.89**	1				
Extrav.	-0.06	-0.07	-0.07	1			
Agreeab.	-0.01	0.08	0.08	-0.01	1		
Consc.	0.17*	0.19*	0.18*	0.24**	0.42**	1	
Neuro.	0.04	-0.02	-0.00	-0.32**	-0.27**	-0.49**	1
Openn.	0.15	0.18*	0.17*	0.28**	0.07	-0.02	-0.13

** Correlation is significant at the 0.01 level (1-tailed).
* Correlation is significant at the 0.05 level (1-tailed).

10.6.4 Hypothesis Testing

The null hypothesis was tested using the one-way analysis of variance (ANOVA) test to analyze whether there was any significant difference in academic performance between the

three levels of Openness to experience (low, medium, and high). ANOVA compares the variance between the groups of low, medium and high Openness and produces the *F* ratio, which represents the variance between the groups. A large *F* ratio indicates that the variation due to the treatment is greater than the variation due to error or unsystematic variation in the data (Pallant, 2007).

Table 10.3 provides the mean and standard deviation values for academic performance for each group. The actual difference in mean assignment scores between the groups was quite small when compared with the other performance measures (midterm test and final exam). Overall mean values indicate that paired students of high Openness performed better in the assignments, midterm-test and exam than the other groups. The results from the Levene's test for homogeneity of variances, shown in Table 10.4, indicate that the variances of scores were significantly different for each group of Openness to experience (i.e. the significance value is less than 0.05). In this case, the homogeneity of variance assumption was violated and therefore instead of referring to the ordinary ANOVA, the *Robust Tests of Equality of Means* needed to be consulted using either the Welch or Brown-Forsythe test (Pallant, 2007).

Table 10.5 shows the output from the Robust Tests of Equality of Means, which provides the adjusted degrees of freedom and the associated *p*-value for the overall ANOVA. Both tests (Welch and Brown-Forsythe) indicate that there was a statistically significant difference between the three levels of Openness to experience relating to the mean scores of paired students' academic performance. Note that the statistic ratio is significant at the 0.05 alpha level. Based on the *p* values, we had evidence to reject the null hypothesis and it can be concluded that at least one of the groups means is significantly different from the others (i.e. $W(2, 87.51) = 4.79$, $p < 0.05$, for assignments; $W(2, 88.81) = 7.43$, $p < 0.05$, for the midterm test, and $W(2, 86.72) = 7.65$, $p < 0.05$, for the final exam).

Table 10.3 Mean and standard deviation of paired students' academic performance

Performance Measures	Openness to Experience Level	N	Mean	SD
Assignments (Range: 0 to 14)	Low Openness	48	11.02	3.58
	Medium Openness	47	10.24	3.41
	High Openness	42	12.06	2.23
	Total	137	11.07	3.23
Midterm Scores (Range: 0 to 100)	Low Openness	48	64.14	22.72
	Medium Openness	47	57.97	26.53
	High Openness	42	75.67	18.89
	Total	137	65.56	24.00
Final Exam (Range: 0 to 100)	Low Openness	48	66.83	26.46
	Medium Openness	45	60.17	28.51
	High Openness	42	79.30	19.35
	Total	135	68.49	26.23

Table 10.4 Levene's Tests

	Levene Statistic (F)	df1	df2	Sig.
Assignments	5.78	2	134	0.004
Midterm Test	5.29	2	134	0.006
Final Exam	6.88	2	132	0.001

Table 10.5 Robust Tests of Equality of Means

Performance Measures		*Statistic	df1	df2	Sig.
Assignments	Welch	4.79	2	87.51	0.01
	Brown-Forsythe	3.79	2	124.13	0.03
Midterm	Welch	7.43	2	88.81	0.001
	Brown-Forsythe	6.79	2	128.02	0.002
Final Exam	Welch	7.65	2	86.72	0.001
	Brown-Forsythe	6.53	2	123.94	0.002

* Asymptotically F distributed

Post-hoc comparisons were performed to further examine which groups means differed (see Table 10.6). For this purpose, we applied the Games-Howell procedure because it was reported to be the appropriate procedure to be used when the assumption of equal variances was violated (Morgan et al., 2004). Table 10.6 shows the multiple comparisons between groups where the difference between group means, the standard error of that difference, the significance level and the 95% confidence interval are displayed for each groups' pair. There was a significant difference in performance between pairs when the significant value was less than 0.05 or whenever both confidence interval were negative. The results from applying the Games-Howell test could be summarized as follows:

1) Paired students of high Openness to experience achieved better performance in assignments, midterm test, and final exam when compared with their counterparts.

2) Paired students of lower and medium Openness to experience had comparable performance in assignments, midterm test, and final exam.

Table 10.6 Post Hoc Test (Multiple Comparison using Games-Howell procedure)

Dependent Variables	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Assign.	Low	Medium	0.779	0.718	0.525	-0.930	2.488
		High	-1.04	0.621	0.221	-2.525	0.443
	Medium	Low	-0.779	0.717	0.525	-2.488	0.930
		High	-1.819*	0.604	0.010	-3.263	-0.377
	High	Low	1.041	0.621	0.221	-0.443	2.525
		Medium	1.819*	0.604	0.010	0.377	3.263
MidTerm	Low	Medium	6.167	5.073	0.447	-5.922	18.256
		High	-11.532*	4.389	0.027	-21.996	-1.069
	Medium	Low	-6.167	5.073	0.447	-18.256	5.922
		High	-17.699*	4.845	0.001	-29.263	-6.136
	High	Low	11.532*	4.389	0.027	1.069	21.996
		Medium	17.699*	4.845	0.001	6.136	29.263
Final Exam	Low	Medium	6.667	5.714	0.476	-6.952	20.286
		High	-12.476*	4.848	0.031	-24.04	-0.912
	Medium	Low	-6.667	5.714	0.476	-20.286	6.952
		High	-19.142*	5.194	0.001	-31.553	-6.733
	High	Low	12.476*	4.848	0.031	0.912	24.04
		Medium	19.142*	5.194	0.001	6.733	31.553

* The mean difference is significant at $\alpha = 0.05$

10.6.5 Statistical Power Analysis

A statistical power represents the likelihood that a treatment effect will be observed whenever there is one. High statistical power indicates greater ability to detect a difference between treatments if a true difference exists, when compared with a study with relatively low statistical

power (Dyba et al., 2006). The post-hoc power analysis herein was conducted using the G*Power (Version 3.1.2) (Faul et al., 2007). Tables 10.7 to 10.9 show the protocols of the power analysis where the input and output parameters were specified. Our analysis indicates that this experiment demonstrates a reasonably high statistical power (between 0.70 and 0.88) with a medium effect size (ranging between 0.24 and 0.30).

Table 10.7 Power analysis protocol (assignments)

F tests – ANOVA: Fixed effects, omnibus, one-way		
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size f	= 0.24
	α err prob	= 0.05
	Total sample size	= 137
	Number of groups	= 3
Output:	Noncentrality parameter λ	= 7.85
	Critical F	= 3.06
	Numerator df	= 2
	Denominator df	= 134
	Power (1-β err prob)	= 0.70

Table 10.8 Power analysis protocol (midterm test)

F tests – ANOVA: Fixed effects, omnibus, one-way		
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size f	= 0.30
	α err prob	= 0.05
	Total sample size	= 137
	Number of groups	= 3
Output:	Noncentrality parameter λ	= 12.32
	Critical F	= 3.06
	Numerator df	= 2
	Denominator df	= 134
	Power (1-β err prob)	= 0.88

Table 10.9 Power analysis protocol (final exam)

F tests – ANOVA: Fixed effects, omnibus, one-way		
Analysis:	Post hoc: Compute achieved power	
Input:	Effect size f	= 0.30
	α err prob	= 0.05
	Total sample size	= 135
	Number of groups	= 3
Output:	Noncentrality parameter λ	= 12.06
	Critical F	= 3.06
	Numerator df	= 2
	Denominator df	= 132
	Power (1-β err prob)	= 0.88

Figure 10.5 illustrates the central (H_0) and the noncentral (H_1) test statistic distributions, the critical *F* value and the associated error probabilities for the power analysis. The red distribution shows the spread of *F*-test values assuming that the null hypothesis was true; the

blue line shows the distribution of *F*-test values when the population effect size was of 0.30. In this experiment, the statistical power of 0.88 indicates that there is only a 12% probability of falsely failing to reject the null hypothesis, thus decreasing the risk of making a Type II error (represented by β).

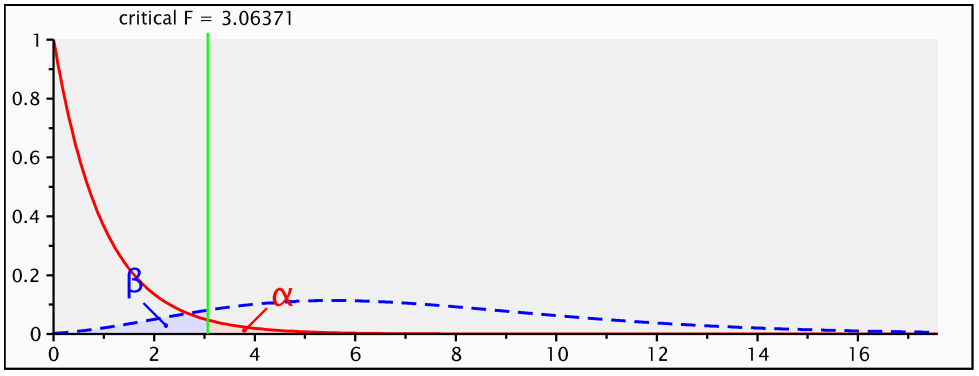


Figure 10.5 Central and noncentral distributions

10.6.6 Results for Satisfaction and Confidence

We analysed paired students' levels of satisfaction and confidence based on data gathered from a PP questionnaire (Appendix B.4) distributed in each tutorial session. Data were gathered for seven weeks of tutorials, starting from the third week until the end of the semester. Data were gathered starting from the third week onwards to give students ample time to familiarize themselves with PP during the first two weeks of tutorials. The questionnaire's response rate was initially 81.7% when gathered for the first time; however it decreased to 56.9% for the last week of tutorials.

Students indicated their level of satisfaction working with their partner by answering the question "*Please rate how satisfied are you working with your partner*", measured on a scale from 1 (very dissatisfied) to 5 (very satisfied). Figure 10.6 shows the participants' responses. On average 75 (87.2%), out of an average of 86 students attending the tutorials, were satisfied working with their partner.

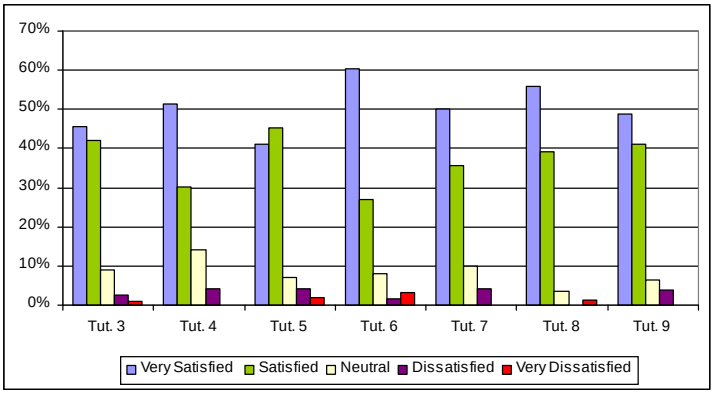


Figure 10.6 Survey on PP satisfaction

The Kruskal-Wallis test was used to compare satisfaction levels between groups of different levels of Openness to experience. Table 10.10 shows the mean satisfaction rank of paired students, where a higher mean rank indicates a higher satisfaction level. The results indicate that there was only one tutorial (i.e. tutorial 4) that showed a significant value (χ^2 (2, 99) = 7.19, p = 0.03); therefore, overall our results demonstrated that the satisfaction level of paired students were not affected by students' level of Openness to experience.

Table 10.10 Mean rank for satisfaction level

	Openness Level	N	Mean Rank	Sig.	Satisfied/ Very Satisfied (%)
Tut. 3 N=112	Low	37	49.27	0.19	87.5
	Medium	37	59.35		
	High	38	60.76		
Tut. 4 N=99	Low	33	48.26	0.03	81.8
	Medium	34	59.16		
	High	32	42.06		
Tut. 5 N=97	Low	35	48.44	0.27	86.6
	Medium	34	54.10		
	High	28	43.50		
Tut. 6 N=63	Low	18	29.31	0.08	87.3
	Medium	31	29.73		
	High	14	40.50		
Tut. 7 N=70	Low	26	34.46	0.84	85.7
	Medium	19	37.58		
	High	25	35.00		
Tut. 8 N=84	Low	28	46.95	0.33	95.2
	Medium	29	38.55		
	High	27	42.13		
Tut. 9 N=78	Low	26	37.48	0.69	89.7
	Medium	24	42.40		
	High	28	38.89		

Students reported their confidence level by answering the question "*How do you rate your level of confidence solving the exercises with your partner?*", measured on a scale from 1 (very low) to 5 (very high). Figure 10.7 shows participants' responses. On average 73 (84.9%), out of an average of 86 students attending tutorials were highly confident in the correctness of their programming solutions when working in pairs. Table 10.11 presents the mean rank for paired students' confidence level, showing only one tutorial with a statistically significant difference in confidence level across the three groups (tutorial 4, χ^2(2, 99) = 8.78, p=0.01). Overall findings indicate that paired students' confidence level was not affected by students' Openness to experience level.

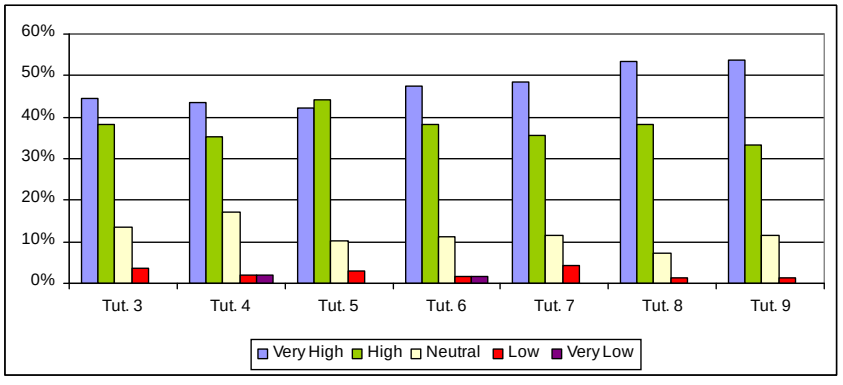


Figure 10.7 Survey on PP confidence

Table 10.11 Mean rank for confidence level

	Openness Level	N	Mean Rank	Sig.	High Confidence (%)
Tut. 3 N=112	Low Medium High	37 37 38	50.15 61.91 57.42	0.24	83
Tut. 4 N=99	Low Medium High	33 34 32	49.02 59.96 40.44	0.01	78.8
Tut. 5 N=97	Low Medium High	35 34 28	43.97 54.22 48.95	0.26	86.6
Tut. 6 N=63	Low Medium High	18 31 14	28.06 31.19 38.86	0.18	85.7
Tut. 7 N=70	Low Medium High	26 19 25	32.98 37.50 36.60	0.68	84.3
Tut. 8 N=84	Low Medium High	28 29 27	47.82 38.03 41.78	0.23	91.6
Tut. 9 N=78	Low Medium High	26 24 28	40.75 40.10 37.82	0.86	87.2

In addition to measuring the satisfaction and confidence level, students' feedback on the following questions were also gathered:

"I felt that working with this partner was a productive experience." (Q1)

"I enjoyed working with my partner." (Q2)

"My motivation level increased when working with my partner." (Q3)

Figures 10.8 to 10.10 show the students' feedback regarding their experience working in pairs. On average 78 out of 87 students (89.7%) indicated that their pairing experience was productive (Q1). In terms of enjoyment, 78 out of 87 students (89.7%) agreed that working with their partner was an enjoyable experience (Q2). PP also helps increased students' motivation level (Q3). On average 73 out of 87 students (83.9%) agreed with the statement mentioned in Q3.

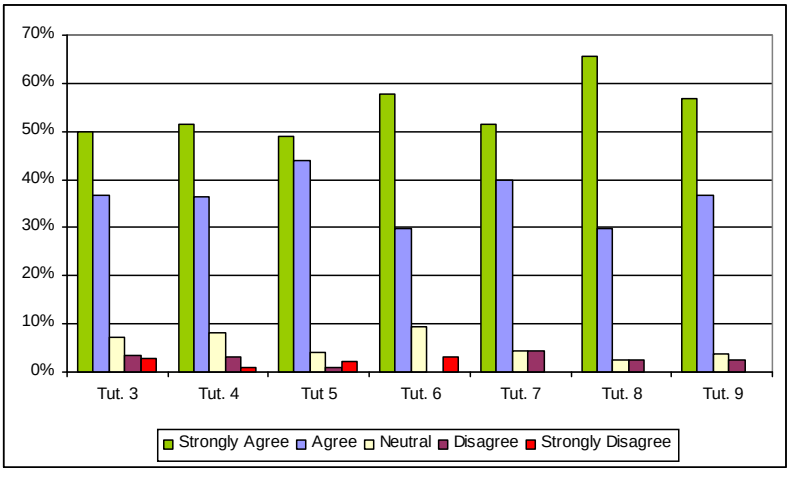


Figure 10.8 Responses on PP's experience (Q1)

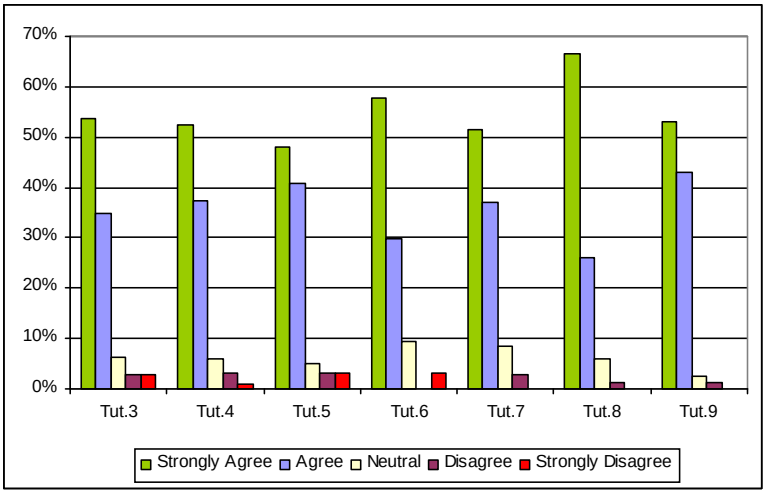


Figure 10.9 Responses on PP's experience (Q2)

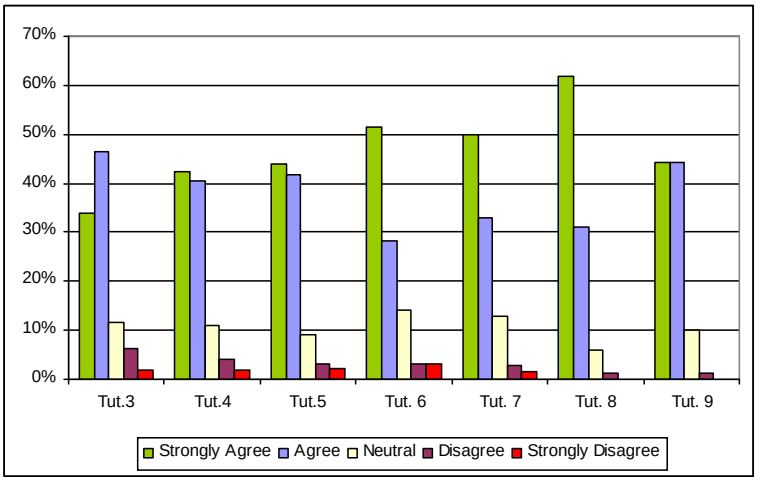


Figure 10.10 Responses on PP's experience (Q3)

10.7 Discussion

The findings from this experiment showed that paired students' academic performance appeared to be significantly affected by students' Openness to experience level. These findings corroborate some existing results reported in the personality-psychology literature. For example, Blickle (1996) found Openness to experience to be positively associated with academic performance. He conducted factor analyses of learning strategies and discovered that the "elaboration" factor for learning strategies was highly correlated with the Openness to experience scale (r = 0.49 in Study 1; r = 0.39 in Study 2). His findings indicate that the Openness to experience trait has a crucial effect on the learning strategies, which mediate the relationship between personality trait and performance (Blickle, 1996).

Paunonen and Ashton (2001) also found that Openness to experience was a significant predictor of academic performance. Their study demonstrated that, in comparison to the broad personality factor (i.e. the Openness to experience), the narrow personality traits of Openness to experience can better predict the academic performance, measured using

students' course grade. For instance, there was a significant positive correlation between the "need for understanding" trait and students' grades (r = 0.23).

Farsides and Woodfield (2003) reported that Openness to experience was positively associated with students' final grades. They mentioned that "*being Open to experience provides academic benefits beyond those provided by being clever and being motivated to turn up to classes*" (p. 1239). Similar findings were also reported in another two studies (Lounsbury et al., 2003; Phillips, Abraham, & Bond, 2003). In another study, Chamorro-Premuzic and Furnham (2008) suggest that students' exam marks are significantly correlated with Openness to experience and deep learning approaches. They also found that Openness to experience mediates the link or relationship between academic performance and IQ, suggesting that high IQ individuals achieve higher grades due to their high level of Openness to experience. It has also been reported that Openness to experience is positively correlated with intelligence within the range of r = 0.20 – 0.40 (Ackerman & Heggestad, 1997).

Ackerman and Heggestad's meta-analysis (1997) also revealed a substantial positive correlation between Openness to experience and intelligence, and "knowledge and achievement". The two major facets of Openness to experience related to lexical intellect are "Aesthetics" and "Ideas" (Johnson, 1994; Saucier 1994). Matzler et al. (2008) have shown in their study that the acquisition and dissemination of knowledge are greater for teams scoring high on Openness to experience.

In the context of paired programming, students working collaboratively in solving programming tasks can benefit from the elements of Openness to experience by being more willing to engage in learning experiences. Studies' findings report that the mean level of Openness to experience in team compositions positively influences knowledge sharing among team members (Hsu et al., 2007; Matzler et al., 2008). It means that a team composed of higher aggregate levels of Openness to experience resulted into higher levels of knowledge sharing (Hsu et al., 2007). LePine (2003) stated that "*In a team setting, open individuals should not only make more suggestions, but because they tend to be insightful, enthusiastic, and talkative, they should tend to build on the ideas of other members*" (p. 32).

It has been suggested that Openness to experience is a better predictor when the situation involves novel or complex tasks (Griffin & Hesketh, 2004). Thus, it is also possible that paired students who are high on Openness to experience were more inquisitive in solving complex issues such as programming problems. This is because open individuals tend to be more creative and receptive to ideas/change and willing to try new thing or learning to do different things (LePine, 2003; Harris, 2004). In our experiment, we found a positive correlation between Openness to experience and paired students' performance in the midterm test and final exam, a result which is consistent with the findings from our previous experiments (see Chapter 6 and Chapter 8). The findings from the present experiment also showed that paired students of high Openness levels outperformed those who have low and medium level Openness, thus confirming our supposition that differences in Openness to experience levels affect the academic performance of students who pair programmed.

10.8 Threats to the Validity

There are several potential threats to the validity of our findings. In this experiment, academic performance was used as our dependent variable and a surrogate measure of PP's effectiveness. However, students' academic performance may also be affected by other factors such as learning styles, self-motivation, and programming ability or competency. In spite of being a surrogate measure, students regularly attend the tutorial and practicing PP throughout an entire semester may have had an influence on their learning process which eventually affected their performance in the test and exam.

Due to the limitation in the sample size employed in this study, we are able to account for only a single personality factor (i.e. Openness to experience) and this prevents us from controlling for the effects of other personality factors towards pairing effectiveness or students' academic performance. For instance, students may perform well or excel in this course because of their conscientious behavior regardless of their high level of Openness to experience. This is evident from the significant positive correlation between academic performance measures (i.e. assignments, midterm, and final exam) and the students' Conscientiousness scores (see Table 10.2). Therefore, we suggest that future replication study should consider controlling the effects of these two major personality factors.

The other methodological limitation of this experiment is that the course was taught by three different instructors and the weekly tutorials were also run by several tutors. Thus, we cannot rule out the possibility that the quality of teaching or the delivery method may have influenced students' ability to comprehend the course.

10.9 Summary

The findings from the present experiment provide strong support to our alternative hypothesis regarding the effects of the Openness to experience factor on paired students' academic performance. We found evidence that the level of Openness to experience played a significant role in influencing students' academic performance where paired students of high Openness achieved better performance compared with their counterparts. The satisfaction and confidence level of students who worked in pairs, however, were not affected by their level of Openness to experience. Results showed that on average 87% of students indicated that their satisfaction level was high when working with their partner. Similarly, most students (85%) responded that they had high level of confidence in solving the programming exercises collaboratively with their partner. In the next chapter, we provide an overall discussion of findings gathered from a series of experiments and the implication of this research for teaching and learning in CS/SE education.

Chapter 11

OVERALL DISCUSSION OF FINDINGS FROM OUR FORMAL EXPERIMENTS

This chapter provides an overall discussion of the findings obtained from the series of formal experiments carried out as part of this research. This discussion includes an analysis and aggregation based on the individual findings from each of the experiments reported in Chapters 6 to 10, an overall discussion about threats to the validity of our findings, and the implications of our research findings for researchers and educators in CS/SE.

11.1 Analysis of Findings

The overall findings discussed herein are based on the results from a series of formal experiments conducted between 2009 and 2010 at The University of Auckland, with participants comprising of first and second year undergraduate students. These experiments were held during tutorials held as part of two CS courses, namely Principles Programming (COMPSCI 101) and Software Design and Construction (COMPSCI 230).

The purpose of the experiments was to investigate the effectiveness of PP as a pedagogical tool for CS/SE education by focusing on the effects of personality traits on the academic performance of students practicing PP. The Five-Factor Personality model was chosen as our personality framework; our research focused on three personality factors of this model, reported as being relevant for a higher education context: *Conscientiousness*, *Neuroticism*, and *Openness to experience*. Table 11.1 summarizes the characteristics of each of the formal experiments we conducted.

Table 11.1 Formal experiments characteristics

Experiment	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
Semester:	Summer 2009	Semester 1, 2009	Semester 1, 2009	Semester 2, 2009	Semester 1, 2010
Sample size:	48	212	77	118	137
Course:	CS101	CS101	CS230	CS101	CS101
Subjects:	First year undergraduate	First year undergraduate	2nd year undergraduate CS students	First year undergraduate	First year undergraduate
Tutorial settings:	• Compulsory • 2 hours • Closed-lab	• Compulsory • 2 hours • Closed-lab	• Optional • An hour • Closed-lab	• Compulsory • 2 hours • Closed-lab	• Compulsory • 2 hours • Closed-lab
Personality factor (IV):	Conscientiousness	Conscientiousness	Conscientiousness	Neuroticism	Openness to Experience

In each experiment, the dependent variables observed were students' academic performance in assignments, in a midterm test, and in the final exam. In addition to investigating the effects of a certain personality factor, the experiments also looked at the

relationship between an individual student's personality score and their academic performance, and the students' level of satisfaction and confidence when working in pairs.

Table 11.2 presents the alternative hypothesis and the results summary for each of the formal experiments aforementioned. In the first experiment (*Exp 1*), we hypothesized that there would be differences on performance between groups of paired students with similar and mixed Conscientiousness. In the second (*Exp 2*) and third experiments (*Exp 3*), we investigated whether different levels of Conscientiousness (low/medium/high) could have had an impact on paired students' academic performance. In Exp 3, we found that results only differed significantly for the midterm test, and these differences were absent for the other dependent variables. Based on the results from these experiments, we could not find supporting evidence for any of the alternative hypotheses. However, the low power level exhibited in these studies suggests that the patterns observed may likely not apply to other samples from the same population of interest.

Table 11.2 Comparison of the five formal experiments (hypothesis & results)

Experiment	**Exp 1 (Chapter 6)**	**Exp 2 (Chapter 7)**	**Exp 3 (Chapter 8)**	**Exp 4 (Chapter 9)**	**Exp 5 (Chapter 10)**
Alternative Hypothesis	Differences in personality trait Conscientiousness affect the effectiveness of students who pair programmed.	Differences in Conscientiousness levels affect the effectiveness of students who pair programmed.	(same as Exp2)	Differences in Neuroticism levels affect the effectiveness of students who pair programmed.	Differences in Openness to experience levels affect the effectiveness of students who pair programmed.
Alternative Hypothesis supported? (Yes/No)	No	No	No (except for the midterm test)	No	Yes
Summary of Results	Lack of evidence for distinguishing performance of paired students between similar and mixed personality group based on *Conscientiousness* trait.	Paired students' academic performance was not affected by differences in *Conscientiousness* levels (low/medium/high).	Except for the midterm test, results showed that paired students' performance was not affected by differences in *Conscientiousness* levels.	Paired students performance was not significantly affected by the different levels of Neuroticism.	Paired students of high *Openness* achieved better performance than the low and medium Openness.

In relation to our fourth experiment (*Exp 4*), we investigated whether differences in levels of Neuroticism (low/medium/high) when pairing had significant impact on students' academic performance. No supporting evidence to the alternative hypothesis was found. Finally, the fifth experiment (*Exp 5*) investigated the effects of Openness to experience on students' academic performance when pairing, and results showed that this factor had a substantial impact towards paired students' performance. In the following subsections we summarize the aggregation of findings in terms of correlations between factors (both IV and DVs) and the overall analysis based on the hypothesis testing of each experiment.

11.1.1 Analysis of Correlation Results

Table 11.3 presents the aggregation of the bivariate Pearson correlation results between the personality factors employed in this research and the corresponding measures of paired students' performance. There was a significant positive correlation between Conscientiousness and paired students' performance in assignments for three of the five experiments, suggesting that the performance in assignments was largely related to how conscientious the students were, and less related to their Neuroticism or Openness to experience levels.

However, students' performances in the midterm test and final exam appeared to be mostly significantly and positively correlated with students' level of Openness to experience. In *Exp 4* and *Exp 5*, Conscientiousness showed a significant positive correlation with most academic performance criteria. Overall, paired students' academic performance was not associated with students' Neuroticism levels.

The results from the correlation analysis indicate that the two personality factors potentially affecting academic performance of paired students in the context of these computer science courses were Conscientiousness and Openness to experience. This is in accordance with some findings reported in the personality-psychology literature conducted within higher academic settings. For instance, Chamorro-Premuzic & Furnham (2008) report that academic performance is positively correlated with Openness and Conscientiousness, and that these personality variables explained approximately 31% of the variance in students' academic performance. Other studies that give support to these personality factors are studies reported by Lounsbury et al. (2003), Dollinger & Orf (1991), and a large-scale meta-analysis of the FFM and academic performance by Poropat (2009).

Table 11.3 Results on correlations (FFM vs academic performance)

Personality Factor	**Exp.**	**Correlation (*r*)**		
		Assign.	MidTerm	Final
Conscientiousness	1*	0.29**	0.07	-0.05
	2*	0.02	-0.07	-0.02
	3*	0.00	0.14	0.09
	4	0.19**	0.19**	0.15
	5	0.17**	0.19**	0.18**
Neuroticism	1	-0.17	-0.04	-0.03
	2	-0.01	-0.06	-0.05
	3	-0.01	-0.01	-0.15
	4*	0.05	-0.01	0.01
	5	0.04	-0.02	-0.00
Openness to Experience	1	-0.05	0.35**	0.29**
	2	0.19**	0.12	0.20**
	3	-0.02	0.25**	0.26**
	4	0.01	0.23**	0.15
	5*	0.15	0.18**	0.17**

N(*Exp 1*) = 48; N(*Exp 2*) = 212; N(*Exp 3*) = 77; N(*Exp 4*) = 118, N(*Exp 5*)=137
(*) Personality factor is controlled
(**) Significant at $\alpha < 0.05$

11.1.2 Analysis of the Hypothesis Testing

Table 11.4 presents the aggregation of the hypothesis testing results and the associated statistical power analysis of each experiment. The results from two experiments involving an introductory programming course (*Exp 1* and *Exp 2*) indicate that there was a lack of evidence to differentiate performance of paired students based on their Conscientiousness levels, whereas in the third experiment (*Exp 3*), which targeted at an advanced level course (Software Design and Construction), a significant finding was only shown for the midterm test (i.e. midterm test scores were affected by Conscientiousness levels). There was also a lack of evidence to support our alternative hypothesis on Neuroticism in *Exp 4*. Finally, we found evidence that supported our alternative hypothesis in *Exp 5* where the Openness to experience trait significantly distinguished academic performance of paired students.

Table 11.4 Hypothesis testing and statistical power

Exp.	N	Personality Factor	Supported Hypothesis? (Yes/No)	Statistical Test (*)	Effect Size	Statistical Power
1	48	Conscientiousness	No	MANOVA	0.08	0.28
2	212	Conscientiousness	No	ANOVA	0.07 (assign.) 0.06 (midterm) 0.04 (final)	0.14 0.10 0.08
3	77	Conscientiousness	No (except for the mid term test)	ANOVA	0.11 (assign.) 0.32 (midterm) 0.10 (final)	0.12 0.71 0.11
4	118	Neuroticism	No	ANOVA	0.20 (final) 0.22 (midterm) 0.18 (final)	0.47 0.54 0.36
5	137	Openness to Experience	Yes	ANOVA	0.24 (final) 0.30 (midterm) 0.30 (final)	0.70 0.88 0.88

(*) Alpha (α) is set to 0.05 in all experiments

Each of the five formal experiments included a post-hoc analysis of statistical power, so to help interpret their results. The statistical power analysis reports the estimated effect sizes and the power level based on the statistical test employed in the experiment. The importance of reporting these data has been emphasized by Dyba et al. (2006) who recommend that "*we should explore in more depth what constitutes meaningful effect sizes within SE research, in order to establish specific SE convention*" (p. 751). In another study, Miller et al. (1997) also stressed that "*Reporting the effect size allows other researchers in the field to judge the importance of the study's results, while at the same time allowing comparison to the findings of previous studies. Moreover, this information will facilitate meta-analyses and cost-effective planning for future research in related areas*" (p. 289).

When investigating for the effects of Conscientiousness, the range of statistical power varied widely from 0.08 to 0.71. These statistical powers were considered to be low compared with the recommended baseline of 0.80, when assessed according to the statistical power's standard convention (Cohen, 1988). Similarly, the range of statistical power when investigating the effects of Neuroticism (i.e. from 0.36 to 0.54) was also below the recommended power. Nevertheless, we observed a sufficient amount of statistical power in

Exp 5 when investigating the effects of Openness to experience on students' performance. The power value indicates a probability of approximately 88% of achieving statistical significance (at $\alpha = 0.05$) in differentiating academic performance between paired students of different levels of Openness to experience (see Table 11.4).

In terms of the effect size, we observed that the effect size also varied widely from 0.04 to 0.32 when differentiating the performance of paired students based on students' Conscientiousness level. These effect sizes were considered to be low to medium size, but in most cases the observed effect size was remarkably low (i.e. between 0.04 and 0.10). These low effect sizes indicate that there was only a trivial impact of the treatment (Conscientiousness) on the dependent variables (i.e. students' academic performance). The range of effect sizes for Neuroticism varied between 0.18 and 0.22, which was nearly a medium effect size according to Cohen's guidelines (Cohen, 1988). Of the three personality factors investigated in this research, the strength of effect for the Openness to experience was found most significant (i.e. medium effect size of 0.24 – 0.30). These effect size indices help in identifying the "practical importance" or meaningfulness of the results (Cohen, 1988; Dyba et al., 2006). Within our context, it represents the magnitude of academic achievement in assignments, midterm test, or exam.

Although many studies support Conscientiousness as the most significant personality factor for predicting academic performance or team's performance (e.g. Dollinger & Orf, 1991; O'Connor & Paunonen, 2007; Poropat, 2009), the results we obtained did not support this view. In our study we could not find significant evidence to distinguish paired students' academic performance based on their Conscientiousness levels. The results from our experiments suggest that the level of paired students' Openness to experience could impact students' academic performance significantly far more than their Conscientiousness. This is in line with studies reported in the personality/educational psychology literature that mentioned the nature or characteristics of open individuals as being bright, broad-minded, and creative, which will eventually bring significant advantage on their academic success (Paunonen & Ashton, 2001; Farsides & Woodfield, 2003; Philips et al., 2003; Lounsbury et al., 2003).

11.1.3 Analysis of Quantitative Surveys

Data on students' feedback about working with their partner was gathered using a questionnaire distributed in every tutorial session. This questionnaire was designed to measure the levels of satisfaction and confidence of paired students. In addition, students also provided feedback on whether the pairing was useful or productive, whether it was an enjoyable experience, and whether or not pairing helped increase their motivation. Table 11.5 presents the summary of results. On average, 86% students gained higher satisfaction from the PP experience and 84% responded that their confidence level in solving the programming exercises were high. Likewise, most students (on average 90%) felt that PP was a productive experience, enjoyable (91%) and helped increase their motivation level (on average 84%).

Table 11.5 Summary of paired students feedback

Item/ Percentage (%)	(% of Agree/Strongly Agree) Exp 1	Exp 2	Exp 3	Exp 4	Exp 5
Satisfaction level	88.5	90.2	79.0	85.7	87.2
Confidence level*	87.9	87.7	75.0	84.3	84.9
Productive Experience	90.4	95.0	86.0	90.0	89.7
Enjoyment	92.6	94.0	89.0	88.5	89.7
Increase Motivation level	86.0	87.0	78.0	84.3	83.9

(*) % indicates responses with High/Very High confidence

11.2 Threats to the Validity of the Findings

This section describes the potential threats that may affect the reliability of our research findings. The threats can be addressed based on four types of validity issues (Cook & Campbell, 1979): statistical conclusion validity, internal validity, construct validity, and external validity.

11.2.1 Statistical Conclusion Validity

Statistical conclusion validity is defined as "*inferences about whether it is reasonable to presume covariation given a specified α level and the obtained variances*" (p. 41, Cook & Campbell, 1979). One of the threats to drawing valid inferences about whether covariation occurs in our sample data relates to the low statistical power obtained from our statistical power analysis. When the level of statistical power is low, the likelihood of making a Type II error increases for the cases where a small sample size was employed and the effect size was relatively small (Murphy & Myors, 2003). Tabachnick and Fidell (2001, p. 329) mentioned that a sample size of at least 20 in each group should ensure "robustness". For the case of our experiment this condition was fulfilled. Therefore this reduces the likelihood of committing a Type II error. The low power observed in some of the experiments indicates that our data do not warrant the conclusion that the population means differ between the studied groups. Therefore, we cannot conclude whether there is any real difference in students' academic performance when paired according to their level of personality trait Conscientiousness, or Neuroticism.

Another threat relates to the violation of assumptions of statistical test used in the experiments. In particular, for the *Exp 5*, the variability of dependent variables' scores for each of the groups was not equal, thus the assumption of homogeneity of variance was violated. Although ANOVA is fairly robust to violation of such an assumption (Pallant, 2007; Morgan et al., 2004), the results should be interpreted with particular caution because in some cases the distribution of scores was highly skewed. In the case where we found that the assumption of equal variances was violated, we applied an appropriate post hoc statistical test such as *Games Howell* as recommended by Morgan et al. (2004).

Regarding the normality assumption of our dependent variables, the ANOVA test requires the distribution of scores to be normally distributed (Pallant, 2007). However, even if the

distribution of scores is not normal, the *central limit theorem* leads us to believe that the sampling distribution of mean scores is approximately normal (Myers & Well, 2003). According to the central limit theorem, mean distributions tend to be close to or approach the normal distribution when the sample size is greater than 5 or 10 per group (Norman, 2010). The ANOVA test is also reported to be fairly robust when the assumption for normal distribution population is not fulfilled (Pallant, 2007; Morgan et al., 2004; Norman, 2010).

11.2.2 Internal Validity

Cook & Campbell (1979) defined internal validity as "*the validity with which statements can be made about whether there is a causal relationship from one variable to another in the form in which the variables were manipulated or measured*" (p. 38). Internal validity threats are related to issues such as experimental procedures, treatments, or background of the participants, of which these issues may affect the validity of the conclusions drawn from the study (Cook & Campbell, 1979). In our experiment, participation was voluntary and therefore we had to rely on personality data only from students who were willing to participate in the experiment by filling out the online IPIP-NEO personality test. This situation can bring bias to our study in particular because the sample could not be considered random. The "self-selected" sampling method used in this study was therefore the main source of threat to the internal validity of the findings.

In terms of the pair configuration process employed during the experiment, the allocation of pairs was done randomly based on students' personality trait levels and the process was automated by the PALLOC software. All participants were first year undergraduate students and their academic background appeared to be generally similar. Therefore, the potential for selection bias was minimized.

There is also a tendency for the results to be biased by the lack of control for gender effects. Earlier meta-analysis suggests that gender may affect personality traits (Feingold, 1994); however secondary analyses by Costa et al. (2001) report that gender differences are small relative to individual variation within a single gender group. More recently, Schmitt (2008) reported an interaction between gender and Neuroticism, and such interaction affected self-efficacy, which in turn affected performance. Our inability to control for gender effects when investigating the effects of personality traits on paired students' performance is due to the limited sample size. Thus, future replication studies should consider gender as a possible factor when investigating the effects of personality traits on PP in order to confirm or refute our findings.

The fact that the courses/tutorials employed in our experiments were taught or handled by several instructors/tutors may introduce an internal threat to the validity of our results. This is because differences in teaching style or delivery method may have had an influence on students' motivation and their comprehension level of the course. Nevertheless, we had the same group of tutors appointed for handling the tutorials in every academic semester included in our experiments, thus allowing us to compare the results across different experiments.

Although students were aware of the experiment's objectives (i.e. from the *Participant Information Sheet* - Appendix B.2) and their own personality traits, they were not aware of the investigated hypothesis. Moreover, upon signing the consent form (see Appendix B.3), students were informed that their participation is voluntary and that their decision whether to participate or not will not affect their grades or relationship with any of the department's members. As researchers we did not have any direct influence on the operation or undertaking of the course. The surveys were also completely monitored by the tutor. These issues reduce the potential for social threats.

11.2.3 Construct Validity

Construct validity is defined as "*the degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations were based*." (Trochim, 2006).

In this research, we have applied the GQM framework to define the research objectives, the specific questions and metrics needed to be measured in the formal experiments (Basili et al., 1999). This helps minimize the potential threat to construct validity through identification of metrics early on during the planning of an experiment.

We also constructed a survey questionnaire intended to measure students' perception regarding their pairing experience in terms of satisfaction, confidence, and enjoyment level while working in pairs. Students' satisfaction was measured based on the "*satisfaction with partner or social aspect*", which is one of the satisfaction types in PP described by Puus et al. (2004). We designed the survey questionnaire so that it was as precise as possible and intended to have students answer all questions. The survey questionnaire was designed using a five-point scale so that subjects can choose the answer that best represents their perceptions of the pairing experience (see Appendix B.4). The surveys were distributed at the end of each tutorial and therefore the time spent on them was quite limited. The results obtained showed that, in most tutorials, students were able to give their responses to most questions.

Another issue relates to the constructs used to represent the dependent variable (i.e. PP's effectiveness). In our experiments, students' individual performance in assignments, a midterm test and final exam were used as surrogate measures of PP's effectiveness. It was reported that a potential drawback of using a surrogate measure is that these do not directly answer the primary question (Whyte, 2006). The measures of academic performance for instance, may also be affected by third party variables such as learning strategies, cognitive ability, self-motivation, or programming competency – those that could threaten the internal validity. However, since the study aimed to improve students' learning due to practicing PP throughout the entire semester, measuring their academic performance is in our view appropriate to be used in our context. Moreover, evidence from our SLR indicates that students' academic performance is one of the metrics categories used by researchers to measure PP's effectiveness (Salleh et al., 2010).

The IPIP-NEO which has been used to measure students' personality profile is a self-report inventory that requires students to give responses on personality items/scales. The main issue with a self-report inventory is the ability of respondents to fake their responses by misrepresenting one's self uncharacteristically; termed as "faking good" or "faking bad" (Johnson, 2005). The tendency for participants to bias their responses commonly occurred in organizational behavior research (Donaldson & Grant-Vallone, 2002). We believe that it is less likely for students to respond in socially desirable ways because their responses will not affect their academic record.

In terms of the validity of the scales used to measure personality, the IPIP-NEO is reported to have good reliabilities against other established personality instruments (e.g. NEO-PI-R) (Johnson, 2005; Goldberg, 1999). The internal consistency reliability estimates (Cronbach's alpha) that of the three personality traits used in this study were 0.81 for Conscientiousness, 0.83 for Neuroticism, and 0.71 for Openness to experience. In order to provide good support for internal consistency reliability, the Cronbach's alpha coefficient of a scale should be positive and usually greater than 7.0 (Morgan et al., 2004; Pallant, 2007)

11.2.4 External Validity

External validity is defined as "*the approximate validity with which conclusions are drawn about the generalizability of a causal relationship to and across populations of persons, settings, and times*" (p. 39, Cook & Campbell, 1979). The purpose of our research was to investigate the effects of personality composition based on the FFM towards PP's effectiveness as a pedagogical tool where the experiments were conducted during closed-lab tutorials monitored by a tutor.

The subjects involved in this research were undergraduate students who enrolled in CS courses and who have worked in pairs when solving programming tasks during the tutorials. Thus, the research results presented herein were applicable or can be generalized within a context of higher education settings in particular CS/SE undergraduate courses/tasks. Nevertheless, four of our formal experiments (Exp 1, Exp 2, Exp3, and Exp 4) presented a low statistical power and this situation reduces the likelihood to scale up the results to a wider population of CS higher education. In the Exp 5 we observed an acceptable level of statistical power in the experiments, thus we had a greater confidence that these results were applicable to a wider context of CS academic settings.

It is important to note, however, that most experiments were conducted using subjects enrolled in an introductory programming course, thus the effects may be different when experimenting using higher or advanced level CS/SE courses in which tasks of greater complexity are carried out. Similarly, our subjects were first year undergraduate students; therefore it might be possible to have different effects when using more mature participants such as graduate or post-graduate students.

11.3 Implications for Research

Based on the outcomes of this research, several implications for research can be drawn. The results of our SLR revealed that there are several factors affecting the PP's effectiveness (Salleh et al., 2010). For example, other than personality factor, other factors include skill level, gender, communication skills, learning-style etc. We found evidence from our SLR that PP works best when students of similar skill levels are paired. Thus, one of the implications for research would be to investigate which factor was the strongest predictor of PP's effectiveness; this can be done by examining regression correlations between these factors.

In our research, academic performance criterions such as assignment, midterm test, and final exam were used as surrogate measures to measure PP's effectiveness. Due to the methodological limitation, we did not measure the effectiveness based on actual performance of students while solving the tasks in pair. For instance, instead of using a surrogate measure, one can gauge pairing effectiveness by evaluating pair performance based on the scores obtained from the exercises solved with the partner or based on the quality of code produced by the pair.

Bowers et al. (2000) mentioned that team performance could be affected by the tasks' level of difficulty. A low difficulty task may intrinsically require fewer cognitive resources of the team and for the case of PP, pairing is reported to be most beneficial when it involves a more complex task (Arisholm et al., 2007). In our study, the senior tutor of COMPSC101 rated the difficulty level of programming exercises for the tutorial as 4 out of 10 using a scale from 1 (*very easy*) to 10 (*very complex*). It is therefore possible that future studies might obtain more significant findings than ours if the tasks employed were of greater complexity than the ones used in our experiment.

In our research, we did not assess variables that could potentially mediate the relationship between personality traits and performance or PP's effectiveness. Research evidence suggests that a team's personality composition may be functional to group processes such as "task cohesion" or "social cohesion" rather than having direct impact on performance (Vianen & Dreu, 2001). With regard to understanding the effects of personality on pair performance, Walle and Hannay (2009) have investigated the nature of PP's collaboration as a mediator variable. Their findings indicate that the impact of personality on pair collaboration might be more significant than the impact on pair performance. Thus future studies may look into possible effects of mediator variables to gain insight into the mechanism underlying the personality-performance relationship in PP.

Another implication of this study would be to investigate the influence of a specific personality facet rather than the broad personality factors. Each personality factor in the Five-Factor Model of personality structure encompasses narrow personality traits, at lower-levels of the personality hierarchy (McCrae & Costa, 1997). For instance, according to the NEO-PI-R personality inventory, Conscientiousness includes facets such as *achievement striving*, *competence*, *deliberation*, *dutifulness*, *order*, and *self-discipline* (Costa & McCrae, 1992a). Research evidence reported by educational psychologists indicates that the narrow

personality facets are generally stronger predictors of academic performance than the broad personality factors (O'Connor & Paunonen, 2007; Paunonen & Ashton, 2001; Chamorro-Premuzic & Furnham, 2003b). For example, *achievement striving* and *self-discipline* have been the strongest and most consistent predictors of academic performance (O'Connor & Paunonen, 2007). Thus, accuracy in performance prediction by personality would increase by employing facets rather than the broader trait; this would help identify which personality facets are highly relevant as the performance determinants of students practicing PP.

Due to a limitation in sample size, each experiment conducted in our study investigated only a single personality factor of the FFM. There is a possibility that students' academic performance may also have been affected by another personality factor in the FFM or other non-personality variables such as intelligence, skill level, or gender. For instance, Nguyen, et al. (2005) report that gender has consistently moderated the personality-academic performance relationship in tertiary education. In another study, interaction between gender and Neuroticism is reported to affect self-efficacy, which in turn, affects performance (Schmitt, 2008). Thus, where possible, a larger sample size should be employed in future study to determine if there is any interaction effect between personality and another factor such as gender which may potentially affect paired students' academic performance.

A further implication is in relation to the issue of statistical power. Based on the post-hoc power analysis of our experiments, we observed a consistently low statistical power in some of the findings. These may be due to the underlying observed effect size which was small. The low statistical power may also result from an inadequate sample size employed in the experiments. It was reported that the statistical power of a study is sensitive towards the sample size (Murphy & Myors, 2003). The effects (or the differences between groups) can be more easily detected when an adequate sample size is employed (Murphy & Myors, 2003; Britt & Weisburd, 2010). Cohen (1992) provides a table that can be used as a rule-of-thumb to obtain a necessary sample size given the significance criterion (alpha) and the effect size value (see Table 11.6). For example, assuming a study which compares means of 3 groups that aims to detect a small effect size, the necessary sample size per group is 322, thus, total sample sizes of 966.

Dyba et al. (2006) have proposed some strategies for increasing statistical power such as: i) *increase the sample size*; ii) *set the significance (alpha) criterion with a more liberal value*; iii) *choose powerful statistical tests*; iv) *reduce measurement error and subject heterogeneity*; v) *obtain balanced group sizes*. It has also been suggested that a study should perform a priori power analysis in order to obtain an estimate of sample size expected to achieve a high statistical power (Britt & Weisburd, 2010; Lan & Lian, 2010).

Table 11.6 N for small, medium, and large effect size at power = 0.80 (Cohen, 1992, p. 158)

	Significance Criterion (α)								
	0.01			**0.05**			**0.10**		
Test	Sm	Med	Lg.	Sm	Med	Lg	Sm.	Med.	Lg.
1. Mean diff.	586	95	38	393	64	26	310	50	20
2. Sig *r*	1163	125	41	783	85	28	617	68	22
3. *r* dif	2339	263	96	1573	177	66	1240	140	52
4. *P* = 0.05	1165	127	44	783	85	30	616	67	23
5. *P* dif	584	93	36	392	63	25	309	49	19
6. χ^2.									
1*df*	1168	130	38	785	87	26	618	69	25
2*df*	1388	154	56	964	107	39	771	86	31
3*df*	1546	172	62	1090	121	44	880	98	35
4*df*	1675	186	67	1194	133	48	968	108	39
5*df*	1787	199	71	1293	143	51	1045	116	42
6*df*	1887	210	75	1362	151	54	1113	124	45
7. ANOVA									
2*g*[a]	586	95	38	393	64	26	310	50	20
3*g*[a]	464	76	30	**322**	52	21	258	41	17
4*g*[a]	388	63	25	274	45	18	221	36	15
5*g*[a]	336	55	22	240	39	16	193	32	13
6*g*[a]	299	49	20	215	35	14	174	28	12
7*g*[a]	271	44	18	195	32	13	159	26	11
8. Mult. *R*									
2*k*[b]	698	97	45	481	67	30			
3*k*[b]	780	108	50	547	76	34			
4*k*[b]	841	118	55	599	84	38			
5*k*[b]	901	126	59	645	91	42			
6*k*[b]	953	134	63	686	97	45			
7*k*[b]	998	141	66	726	102	48			
8*k*[b]	1039	147	69	757	107	50			

Note: ES = population effect size, Sm = small; Med. = Medium; Lg. = Large, diff = difference
[a] Number of groups. [b] Number of Independent variables

11.4 Implications for CS/SE Educators

The findings from our study imply that pairing students according to either Conscientiousness or Neuroticism levels do not appear to be significant in affecting paired students' academic performance in CS undergraduate courses. However, due to the low statistical power observed from our analysis, it is inappropriate to suggest that our results would correspond to what would be most likely to occur in a higher education environment when students are paired based on their Conscientiousness or Neuroticism level. There is a possibility that Conscientiousness and Neuroticism may give a significant impact in future replication studies conducted under similar experimental settings (i.e. similar courses and subjects' background). In light of this, our empirical evidence showed mixed findings with regard to the impact of Conscientiousness levels on the academic performance of paired students attending a more advanced computing course. Thus, it would be necessary to conduct a further study involving more complex tasks to determine whether task difficulty level plays a significant role in differentiating paired students' academic performance based upon their personality traits.

Of the three personality traits investigated in our study, we found evidence that paired students' academic performance is significantly affected by their Openness to experience levels. Our results showed a greater performance of high Openness students than those of lower Openness levels. Farsides and Woodfield (2003) note that Openness to experience is highly relevant for educational settings that promote and reward critical and original thought.

Thus, we believe that these results may indicate that this trait may probably be the most important or significant for the development of academic success of CS/SE students. Future replication studies are needed to help strengthen the evidence obtained from our study. It may also be useful to conduct a study that investigates the impact of the two other FFM's personality factors (i.e. Agreeableness and Extraversion) in relation to PP's effectiveness.

One of the practical implications of this study is that PP does not appear to give harmful effects of either students' satisfaction or confidence level in an introductory learning to program course. The results from our studies indicate that students' motivation, enjoyment, satisfaction, and confidence level when working in pairs were very encouraging regardless of their differences in personality trait profiles. Our results were consistent with those existing findings reported in the PP literature (DeClue, 2003; Hanks, 2006; Mendes et al., 2005). This result should support educators in continuing to employ PP as a pedagogical tool in an introductory learning to program course.

The findings obtained from our study were applicable within the context of undergraduate students' learning in an introductory programming course. Thus, further research is needed to extend this study whether findings converge or diverge when employing senior level students such as graduate or postgraduate students. In addition, performing a qualitative study in the future may be practical in order to better understand the results obtained in the present study.

11.5 Summary

The findings from our study provided empirical evidence on the effects of personality traits Conscientiousness, Neuroticism, and Openness to experience in differentiating academic performance of paired students. We found evidence that Openness to experience had significant impact on paired students' academic performance. However, with regards to the other two traits (i.e. Conscientiousness and Neuroticism), no supporting evidence was obtained showing their significance to improve PP's effectiveness. In this chapter, the threats to the validity of our findings were discussed and implications for further research and practice for CS/SE educators were also highlighted. Some of these implications include a proposition for pair formation based on personality traits, and suggestions for future research to further extend our experiments.

Chapter 12

CONCLUSIONS AND FUTURE WORK

This chapter concludes the research described in this thesis by presenting a summary of the research conducted, its contributions and limitations, and some recommendations for future work. Some final remarks about our research are also given.

12.1 Research Summary

The research described in this thesis was undertaken to investigate the effect of personality traits towards the successful implementation of PP in a higher education setting. The research motivation is driven by the evidence from our Systematic Literature Review (SLR) results that discovered the inconsistencies in findings from the PP literature regarding the effects of personality on PP's effectiveness (Salleh et al., 2010). Due to a wide variety of criticism on the use of the Myers-Briggs Type Indicator (MBTI), which has been used in most existing PP research in higher academic settings, we decided to apply the big-five or five-factor personality model (FFM) in this research to measure personality traits. The FFM personality trait model was chosen due to its growing acceptance by the psychological scientific community as a comprehensive taxonomy of human personality (Barrick et al., 1998; Burch & Anderson, 2008).

A series of formal experiments were then conducted at the University of Auckland from 2009 to 2010 to examine the effects of personality on paired students' academic performance. In total, five formal experiments were carried out, where the experimental subjects were undergraduate CS students who volunteered to participate. Three of those five experiments investigated the effects of the personality trait Conscientiousness on the academic performance of paired students attending either a first year introductory programming course or a second year software design and construction course. The fourth and fifth formal experiments investigated the effects of the personality traits Neuroticism and Openness to experience, respectively, where experimental subjects were all first year undergraduate CS students who volunteered to participate. The choice of personality traits was motivated by existing literature where they are generally reported to be educationally important and relevant for higher education (De Raad & Schouwenburg, 1996; Blickle, 1996).

Based on the results from our five formal experiments, we found evidence suggesting that the personality trait Conscientiousness would not affect the academic success of paired students in CS courses. These results were counterintuitive to many findings reported in the educational-psychology literature which suggest Conscientiousness as being significantly related to students' academic performance (Poropat, 2009; Busato et al., 2000). However it is important to highlight that given the statistical power analysis in this research presenting a

lack sufficient statistical power to detect effects of interest, we could not generalize our results to a wider CS/SE population.

Similar to Conscientiousness, differences in Neuroticism levels (low/medium/high) were found not to significantly affect paired students' academic performance; however, once again the low statistical power obtained from our analysis refrain us from concluding the effects of this personality trait on students' academic performance. These results mean that the possible effects of Conscientiousness and Neuroticism may not be ruled out completely. Under such a low statistical power, a positive finding might be obtained in a future study if an adequate sample size and/or a more sensitive research design are employed in the study.

Conversely, Openness to experience showed a statistically significant effect on academic performance where paired students consisting of high Openness achieved better academic performance compared with their counterparts. In this case, our statistical power analysis showed that we had an approximately 88% probability of correctly rejecting the null hypothesis if it is false. Our findings also indicate that despite the variation in students' personality profile when pairing, PP not only caused an increase in satisfaction and confidence level, but also brought enjoyment to the class and helped enhance students' learning motivation.

12.2 Research Contributions

This research has made several contributions to the body of knowledge in the domains of Software Engineering and Computer Science education, summarized as follows:

- The Systematic Literature Review (SLR) presented in Chapter 2 provides the *state-of-the-art* of PP research conducted within a higher education setting. The SLR comprised of 9 years of PP research (1999 – 2007) accumulating and aggregating evidence regarding PP's effectiveness within the higher education context and its potential to be used as a CS/SE pedagogical tool. The major contribution in the review work was the SLR's synthesis of evidence that has identified factors potentially affecting PP's effectiveness for CS/SE students. It also identified measurement methods of PP's effectiveness and the associated quality attributes. Our meta-analysis revealed that PP is effective in improving students' grades on assignments. The SLR uncovered that practicing PP does not bring any detrimental effect to the students' learning and is regarded as beneficial for improving students' learning outcome (Salleh et al., 2010). The outcomes of our SLR can better inform educators wanting to incorporate PP into a CS/SE curriculum and provide implications for research and practice. These include the need to replicate PP studies in areas where findings were inconsistent, or to conduct studies in areas where there is scarcity of or no evidence regarding the effect of certain factors towards PP's effectiveness as a pedagogical tool.

- Our research includes an additional review of literature from other domains such as educational and personality-psychology which provide a foundation for understanding the major personality theories. This also assists in the development of research programs relating personality with academic performance as well as in an effective personality team composition. The knowledge in these areas is preliminary for addressing the role of personality within the CS/SE education context, and in particular to gauge the potential influences on PP's effectiveness. The review also helps in setting out the motivation for selection of the personality framework and personality instrument used in our research (i.e. the FFM and the IPIP-NEO). Of the five personality frameworks discussed (i.e. FFM, MBTI, Keirsey Temperament Sorter, Cattel's 16 Personality Factor, and Eysenck Personality), the FFM was the most notable taxonomy of personality that receive the most support by personality-psychologists (Barrick et al., 1998; Burch & Anderson, 2008; Digman, 1990).
- The first three formal experiments carried out as part of this research focused on understanding the effects of the personality trait Conscientiousness on PP's effectiveness. Contrary to a commonly held view in the educational and personality-psychology literature, our results showed that paired students' academic performance was not significantly affected by their Conscientiousness levels. However, the low statistical power obtained from our analysis limits our ability to generalize this finding into a wider CS/SE population. Although we have seen some significant positive correlations between Conscientiousness and academic performance, our data did not show any significant differences in performance between paired students of different Conscientiousness levels. Future study may replicate our formal experiment in order to confirm or refute our findings. In addition, it is probably worthwhile to consider studying the existence of moderator variables that may affect the personality-performance relationship in PP and the influence of task's complexity (Walle & Hannay, 2009).
- The fourth formal experiment investigated the effects of Neuroticism on paired students' academic performance. Similar to Conscientiousness, our results showed that paired students performance was not significantly affected by the different levels of Neuroticism for the sample employed in this study. The lack of support for the alternative hypothesis could be attributed to the low complexity of the task assigned to the students, and perhaps the existence of moderator variables mediating the relationship between personality trait and performance. Regardless of any possible threats to the validity of the results, the lack of statistical significance might have been due to the lower statistical power observed in this experiment; hence limit the external validity of this finding. Increasing sample size in future replication study may help increase the statistical power.

- The fifth formal experiment carried out showed a significant effect of the personality trait Openness to experience in differentiating paired students' academic performance in assignments, midterm test, and examination. A reasonably high statistical power (between 0.70 and 0.88) demonstrated in this experiment with a moderate effect size estimate (ranging between 0.24 and 0.30) gave us greater confidence that this personality trait had a significant influence on students' academic performance. Our evidence suggests that this particular trait had a significant impact on CS/SE students' academic success when applied on an introductory programming course. We have also discussed the possible reasons behind the results obtained from our experiment, which among them include the type of learning strategies typically employed by students of higher Openness to experience, and that Openness to experience is reported to be associated with intelligence and knowledge sharing (Chamorro-Premuzic & Furnham, 2003; Matzler et al., 2008).

In summary, the results from our SLR showed that students practicing PP achieve productivity similar to or better than solo students; hence PP has the potential to be beneficial for improving students' learning when applied as a pedagogical tool (Salleh et al., 2010). We suggest that educators willing to use PP in their classroom should pair students according to their skill level to achieve greater pair compatibility. The findings from our quantitative surveys in tutorials showed that most students perceived higher satisfaction and confidence from the PP experience (86% and 84%, respectively); approximately 91% perceived it as enjoyable, and 84% responded that working in pairs helped increased their learning motivation. The results from our formal experiment can inform educators that pairing students of high Openness to experience level could be useful for achieving better academic performance. We found no significant evidence regarding the impact of either Conscientiousness or Neuroticism on students' academic performance in an introductory CS programming course.

12.3 Limitations

Except for the third formal experiment, all of our experiments investigated the effects of personality on academic performance within the context of an introductory programming course, thus limiting the generalization of our results to a wider context (e.g. second and third year CS/SE courses). Given that a task' complexity may play a role in affecting the personality-performance relationship (Arisholm et al., 2007), further work in this area needs to be conducted given that the complexity of tasks is very likely to increase in more advanced level courses.

Due to a methodological limitation, the research did not assess the actual performance when students worked in pairs. Rather, performance was measured based on an individual student's achievement in the course using assignments, midterm test, and exam scores. This constraint may have biased the results due to external confounding factors which were not controlled for in the experiments (e.g. cognitive skills, learning strategy, self-motivation, self-esteem).

Another limitation relates to the fact that each of the five experiments investigated only a single personality factor of the FFM due to some constraints such as sample size. Therefore there is a possibility that results obtained from these experiments may also have been affected by existing interactions between some of the FFM personality factors. However, in order to include all the five personality factors within a single experiment (even only two factors) a considerably larger sample size would have been needed.

The personality instrument used in our research (IPIP-NEO) is a self-report inventory that requires students to answer individually the questions pertaining to their behavior/responses on certain situations. The major issue with the use of a self-reporting inventory is the tendency for participants to bias their responses (Donaldson & Grant-Vallone, 2002). This is due to the reason that people tend to respond in socially desirable ways. In order to avoid or minimize bias, using multiple sources of data is reported to be a desirable strategy (Donaldson & Grant-Vallone, 2002).

12.4 Future Work

As previously stated, most of our formal experiments were conducted in an introductory CS programming course, where the complexity of tasks is likely to be much lower than in second or third year CS/SE courses. Therefore, we believe that future work should investigate whether the personality traits of pairs actually do impact upon the performance of design/testing tasks or tasks of higher difficulty level than those from an introductory CS programming course. This would contribute to the PP and CS/SE education bodies of knowledge by also investigating the effect that task type or task complexity has on performance of paired students.

In relation to this issue, in one of our experiments which involved a software design course, we obtained mixed results regarding the effects of the personality trait Conscientiousness on paired students' academic performance (i.e. results are significant only for one dependant variable (DV), but insignificant for other DVs). We identified that one of the limitations of this particular experiment was related to its sample size. Therefore one possible avenue for future work area is to replicate this experiment with a larger sample.

In all of the five formal experiments conducted as part of this research, the effects of personality traits were investigated from the perspective of a broader-level or higher level personality trait. Future work could comprise studies that would assess the effects or influence of personality facets, also known as lower-level traits, in order to establish a greater degree of accuracy in terms of how personality traits can affect paired students' performance. Burch & Anderson (2008) in their review of the state of the science in personality suggest that "*research at facet level can make a useful contribution to our understanding of personality at work*" (p. 285). Therefore, future studies should give attention to the use of narrower traits.

We also suggest that as part of future work it would be interesting to examine the existence of moderator variables that could potentially mediate the personality-performance relationship in PP. We believe that considering mediating processes is important to the extent

that they provide insight into how a certain personality factor affects students' performance. For instance, Walle & Hannay (2009) investigated "pair collaboration" as a mediator variable in a personality-pair performance relationship using professional programmers, and their initial results suggest that "*personality might affect pair collaboration, and that the impact of personality on pair collaboration may be more visible than the impact on pair performance*" (p. 212). Similar studies should be replicated using students in tertiary institutions as subjects.

Further research might explore the issue of personality in PP using a qualitative approach such as case study, ethnography, grounded theory or content analysis. We believe that qualitative investigation may facilitate in further deepening our understanding of the research results in the sense that it helps to discover the nature of pair collaboration from the perspective of FFM traits (Walle & Hannay, 2009). Qualitative studies typically collect various forms of data and portray the issue in its multifaceted form, which we believe would help increase our understanding of the personality traits phenomenon in PP.

Other suggestions for future work include investigating gender-related issues. Evidence from our SLR indicates that there are very few studies looking at the effects of gender and its relation to improve PP's effectiveness in higher education settings (Salleh et al., 2010). Thus, the issue of whether or not pairing students by gender (i.e. female pairs, male pairs or mixed pairs) is beneficial for students' performance is still not clearly understood. In addition, gender differences in personality are reported to appear significant for some personality traits; in particular, women consistently scored higher in Neuroticism and Agreeableness facets (Costa et al., 2001). Thus, future work investigating the effects of these personality factors on PP's effectiveness may consider manipulating the gender variable.

Another aspect that we believe is an interesting direction for future work relates to exploring whether PP mitigates Neuroticism, at least for students engaging in PP tasks. This is due to the fact that Neuroticism denotes a lack of emotional stability. Therefore, it certainly seems plausible to examine whether or not the pairing work facilitates high Neuroticism students to better cope with anxiety, depression, and other negative aspects of Neuroticism.

The two other personality factors that have not been addressed in this study also merit investigation. First, Agreeableness which relates to the degree of friendliness, tolerance, helpfulness and straightforwardness, may have a tendency to influence pair compatibility. A pair comprising of a student who is less tolerant, less considerate, or less friendly may intimidate his/her partner. Second, Extraversion, which indicates the level of talkativeness, enthusiasm, and assertiveness, also potentially affects pair's effectiveness when working together. Having an extravert partner may be helpful in terms of having a stimulating discussion and increasing amount of communications within pairs. Nevertheless, highly extravert pairs may suffer negative consequences of having task disruption by higher levels of interaction. A regression study by Hannay et al. (2010) involving 196 professional software developers discovered Extraversion as the strongest predictor of pair performance. In another correlation study, Extraversion is positively correlated with software quality and software teams with a higher aggregate on Agreeableness achieved the highest job satisfaction

(Acuna et al., 2009). Therefore, research into these personality factors may result in a better understanding of their influence on PP's effectiveness in higher education settings.

12.5 Final Remarks

This research presents empirical evidence regarding the effects of the FFM's personality traits towards improving PP's effectiveness as a CS/SE pedagogical tool. The results showed that the impact of personality trait Conscientiousness and Neuroticism in isolation appear to be insignificant for distinguishing paired students' academic performance in CS courses, at least based on the sample data employed in our research. However, in general these findings were considered inconclusive given the low statistical power to detect the treatment variance in our experiments. The poor power probably results from the small sample size employed in our experiments and/or the small effect sizes observed. On the other hand, levels of Openness to experience that paired students have does indicate their ability to excel in learning within a programming course. Our data showed evidence that the strength of effect for this personality trait was significant with estimated effect size ranging between 0.24 and 0.30; hence indicates its practical significance or importance for distinguishing students' academic performance. These findings shed some light on our understanding of the influence of personality traits in PP from the perspective of FFM.

Personality, by its very nature, is a complex combination of traits or individual characteristics that strongly influence the way people perceive and behave or react towards a certain situation. Understanding how personality traits may assist students in performing well in academic studies should not be neglected, in particular in our attempt to improve the practice of PP as an effective pedagogical tool in higher education institutions.

APPENDIX A: SLR RESOURCES

A.1 Protocol for Systematic Review of Pair Programming Studies in Higher Education Settings

1. Background

In recent years, pair programming technique has been widely adopted in Computer Science/Software Engineering (CS/SE) education in higher education institutions (Preston, 2006). Being popularised initially through the Extreme Programming (XP) methodology, pair programming had significantly paved the way for the vast amount of research to determine its usefulness and effectiveness as a CS pedagogical tool (Mc Dowell et al., 2003; Nagappan et al., 2003; Slaten et al, 2005).

Williams et al. (2000), defined pair programming as a practice in which two programmers sitting side-by-side using only one computer to work collaboratively on the design, algorithm, code or test. The pair consists of two developers who change their role alternately as the "driver" and "navigator". The "driver" is responsible for typing the code and has control over the resources such as computer, mouse and keyboard, whereas the "navigator" or "reviewer" has the responsibility of observing the driver's work.

Early research on pair programming had mainly focused on the ability of the technique to benefit the students in terms of productivity and quality of work produced (Williams et al, 2000). Besides that, research evidence also suggested that pair programming can cause enjoyment (Mc Dowell et al, 2003; Mendes et al., 2005; Williams et al., 2000); increase student's confidence level (C. McDowell et al. (2003 & 2006)); reduce staff workload (Cliburn, 2003); improve course completion rate (C. McDowell et al., 2003; Nagappan et al., 2003); improve performance on exams (Mendes et al., 2005; Nagappan et al., 2003) and increase efficiency in helping female students to work in programming tasks (Berensen et al., 2004; Werner et al.,2004).

Research results however are found to be contradictory as highlighted by Gallis et al (2003). The claims made by several authors regarding the benefits of the practice were argued with the costs (number of hours) incurred from its implementation. For instance, one study conducted by Nawrocki and Wojciechowski (2001) found that there is almost no difference in the development time between the study groups (pair and non-pair). They suggested that the pairing practice would instead double the costs, where the quality of codes remained similar between those groups.

Overall, many studies conclude that students typically had a very positive attitude towards the pair programming practice (Williams et al, 2003; E.A. Chaparro, 2005; Cliburn, 2003; Nagappan et al., 2003; Howard 2006). The potential of effectively using the technique is said to be highly connected with the compatibility factors among the subjects (E.A. Chaparro, 2005). This has been the major issue raised by Katira et al. (2004) where they proposed that compatibility of pair programmers has significant impact on the work productivity. Precisely, they focused their investigation on major factors that determine the compatibility of the pair programmer such as personality type, perceived skill level, perceived technical competence and self-esteem as. In 2006, Williams et al. had performed a similar study to determine whether or not a course instructor can proactively form compatible pairs based on personality type, learning style, skill level, self-esteem, work ethic and time management preference. Their findings showed that pair compatibility can still be achieved based on random pairing, without necessarily considering students' personality type, skill level, self-esteem, work ethic or time management skills (Williams et al., 2006).

Considering the importance of identifying and/or understanding factors potentially contribute to the effectiveness of pair programming as a pedagogical tool, a systematic review (SR) need to be held to assess the availability of existing PP empirical research conducted within higher educational institutions. The SR could further suggest gap(s) or important area

of research in future studies. Therefore, this protocol is developed as a framework to conduct the SR based on the procedures of Kitchenham & Charters (2007).

2. Research questions

Primary Question
What evidence is there of PP studies conducted in higher education settings that investigated PP's effectiveness and/or pair compatibility for CS/SE education?

Structured Questions
The formulation of research question(s) involves four major components (ie. PICOC): Population, intervention, comparison and outcomes (Petticrew & Roberts, 2006). Our primary research question can be decomposed into the following sub-questions:

Question 1
What evidence is there regarding compatibility factors that affect pair compatibility and/or PP's effectiveness as a CS/SE pedagogical tool?

Question 2
Which pairing configurations are considered as most effective looking at the compatibility factors obtained from Question 1?

Question 3
How was PP's effectiveness measured in PP studies and how effective has PP been when used within higher education settings?

Question 4
How was quality measured in the PP studies that used software quality as a measure of effectiveness?

Table 1 shows the summary information on population, intervention, comparison, outcomes, and context involved in the SR process.

Table 7: Summary of PICOC

Population	Computer Science/Software Engineering students
Intervention	Pair Programming
Comparison	None
Outcomes	PP effectiveness
Context	Within the domain of CS/SE teaching and learning addressing effective pairing.

3. Identifying relevant literature

The process of identifying relevant literature involves a comprehensive and exhaustive searching of studies to be included in the review Kitchenham & Charters (2007). This includes the strategy to derive relevant search terms to be used during the search process. The identification of sources primarily from online databases, journals, conferences and grey literature is important to be identified to ensure the wide coverage of potential sources.

3.1 Strategy used to derive search terms

The strategy used to construct search terms is as follows:

a) Major terms can be derived from the review questions based on the population, intervention and outcome (See Table 2);
b) List down all keywords mentioned in the articles. (See Table 3)

c) Other search terms can also be identified looking at the synonyms or alternative words. The words can be searched from the Words Thesaurus function. Content expert, subject librarian or information specialist should also be consulted for further advice in the proper use of the terms. (See Table 4)
d) Use the Boolean OR to incorporate alternative spellings and synonyms (See Table 5)
e) Use the Boolean AND to link the major terms from population, intervention and outcome (See Table 6).

NOTE: Whenever a database does not allow the use of complex Boolean search strings we will design different search strings for each of these databases. The search strings will be piloted and the results of the pilot will be recorded.

Table 8: Terms derived from PICOC

Population	CS/SE Students
Interventions	Pair programming
Comparisons	N/A
Outcomes	Effectiveness
Context	CS/SE/IT education

Table 9: Terms derived from keywords found in papers (sorted descending on year)

Author(s)	**Year**	**Keywords**	**Index Terms/ General Terms**
Mendes et al.	2006	Pair programming, collaboration, software design	Experimentation, measurement
Hanks	2006	Pair programming, student attitudes, student confidence, instructor influence, empirical software engineering, computer science education	G. Terms: Experimentation, Measurement
Muller	2006	Pair Programming, preliminary studies, post-development test-cases	-
David Preston	2006	IT Educational Research, collaborative learning, cooperative learning, pair programming, pedagogy	Management
Katira et al.	2005	Pair programming, compatibility, programming teams	Management, Human Factors
Muller	2005	Pair programming, peer reviews, empirical software engineering, controlled experiment	-
Mendes et al	2005	Pedagogy	Experimentation, Human Factors
Werner et al.	2004	Pair Programming, collaboration, gender	Experimentation, Human Factors
Nagappan et al	2003	Pair programming, collaborative environment, Computer Science education	-
Thomas et al.	2003	Pair programming, self-confidence, first year programming, CS1, closed Labs	Human Factors
Williams et al.	2002	Pair programming, collaborative learning, Computer Science education, Extreme Programming, XP	-
Cockburn et al.	2001	Pair programming, collaborative programming, extreme programming, code reviews, people factors	-
Williams et al.	2000	Pair-programming, collaborative programming, productivity, quality	-

Table 10: Terms derived based on synonym words

Basic terms	Alternative terms
Student	Undergraduate
Pair programming	Pair-programming (Some papers use hyphen)
experiment	Measurement, Evaluation, assessment
Effectiveness	Efficient, successful

Table 11: Concatenation of alternative words using Boolean OR

No.	Results
1	(Student OR undergraduate)
2	(Pair programming OR Pair-programming)
3	(Experiment OR Measurement OR evaluation OR assessment)
4	(Effectiveness OR efficient OR successful)

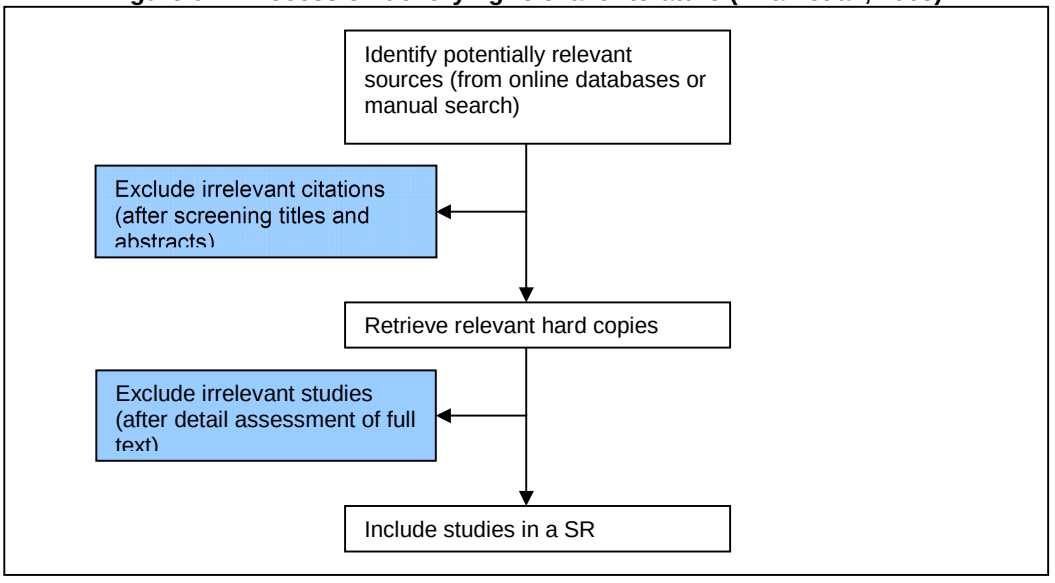
Table 12: Concatenation of all possible words using Boolean AND

Results
(student OR undergraduate) AND (pair programming OR pair-programming) AND (compatibility OR Personality type OR Ethnic OR Self-esteem OR Confidence OR Gender OR skills) AND (Experiment OR Measurement OR evaluation OR assessment) AND (Effectiveness OR efficient OR successful)

4. Searching stages

The process of identifying relevant literature involves several stages. It should be comprehensive and iterative in order to capture as many relevant studies as possible. Figure 1 shows the process involved in identifying relevant literature for a SR (Khan et al., 2003). The searching of literature will cover the study published within the period of 1999 to 2007. We will also focus on papers written only in English.

Figure 0.1: Process of identifying relevant literature (Khan et al., 2003)



4.1 The Primary Search Phase

The initial phase of our search process involves identifying candidate primary sources based on our knowledge on existing PP studies and list of electronic databases subscribed by The University of Auckland. The primary searches will be based on:

Online Databases:

- ACM Digital library
- Current Contents
- EBSCOhost
- IEEExplore
- ISI Web of Science
- INSPEC
- ISI Proceedings
- ProQuest
- ProQuest Dissertations & Theses
- Sage Full Text Collections
- Science Direct
- SpringerLink
- Scopus

Online Search Engine:

- Google Scholar
- CiteSeer
- Agile alliance

4.2 The Secondary Search Phase

The second phase of the search process will support the electronic search activity by reviewing all reference lists in the papers identified from the primary sources. If a paper is found relevant for the SR, it will be added in the existing list of studies to be included in the SR.

5. Documentation of References

Bibliography Management

All references will be stored using the EndNote software. The list of references in the EndNote is known as a *library*. (Refer to ***My Documents/My References/PP.enl***). We will utilize the features of "Direct Export" and "Import Filters" from electronic databases to automatically import all relevant references based on the search string. Information will be entered manually for databases that do not support EndNote. The citations found in each search will be checked against existing references in our library in order to avoid duplication. This is done based on the information on author(s), year and the title.

6. Quality assessment criteria

Inclusion and exclusion criteria

Our inclusion criteria aimed to only include studies that investigate the effectiveness of pair programming in higher educational institutions involving students as subjects. In addition, studies which focused on factors affecting PP's effectiveness, and/or measuring effectiveness and associated quality attributes will also be included.
In terms of exclusion criteria, we will exclude studies relating to the following aspects:

- Papers presenting unsubstantiated claims by the author(s) with no supporting evidence.
- Papers about Agile/XP describing development practices other than PP, such as test-first programming, refactoring etc.
- Papers that only described tools (i.e. software or hardware) that could support PP.
- Papers involving students but outside higher education.
- Papers that solely investigated distributed PP.

- Papers not written in English

Preliminary selection process

During the initial selection process, we will perform screening on the titles and abstracts to see the relevance of the sources. The principal researcher and her supervisor will read through a reference list which contains authors' information, the titles and the abstracts, for the purpose to decide whether to include or exclude a study. Full papers will be obtained whenever they meet the minimum requirement of the inclusion criteria. Full text article will also be referred whenever decision cannot be made based on the title and abstract of a paper.

Final selection process

During the final selection process, the principal researcher will review paper details. During this phase, hard copies of the selected paper will be obtained and full-text article will be referred.

7. Study quality assessment checklists

In assessing the quality of studies, we developed a checklist consisting of questions pertaining to the quality aspect of an article (see Table 13). The following checklist was designed based on the questions proposed in Leedy & Ormrod (2005), Fink (2005), Greenhalgh (2000), Spencer et al. (2000), and Petticrew & Roberts (2006):

Table 13: Study Quality Checklist

No	Item	Answer
1	Was the article referred?	Yes/No
2	Were the aim(s) of the study clearly stated?	Yes/No/Partially
3	Were the study participants or observational units adequately described? For example, SE experience, type (student, practitioner, consultant)	Yes/No/Partially
4	Were the data collection carried out very well? (*)	Yes/No/Partially
5	Were the potential confounders adequately controlled for in the analysis?	Yes/No/Partially
6	Were the approach to and formulation of the analysis was well conveyed? (**)	Yes/No/Partially
7	Were the findings credible? (***)	Yes/No/Partially

Notes/Guidelines:

(*) Quality indicators (i.e possible features for consideration) are as follows:

- Discussion of:
  - Who conducted data collection
  - Procedures/documents used for collection/recording
  - Checks on origin/status/authorship of documents
- Audio or video recording of interviews/discussions/conversations (if not recorded were justifiable reason given?)
- Descriptions of conventions of taking notes (e.g to identify what form of observations were required/to distinguish description from commentary/ analysis)
- Discussion of how field work methods or setting may have influenced data collected
- Demonstration, through portrayal and use of data, that depth, detail and richness were achieved in collection

(**) Quality indicators are as follows:

- Descriptions of form of original data (e.g use of verbatim transcript, observation or interview notes, documents etc)
- Clear rationale for choice of data management method/tool/package

- Evidence of how descriptive analytic categories, classes, label etc (ie. Either through explicit discussion or portrayal in the commentary)
- Discussion, with examples, of how any constructed analytic concepts/typologies etc. have been devised and applied

(***) Quality indicators are as follows:

- Findings/conclusions are supported by data/study evidence (i.e the reader can see how the researcher arrived at his/her conclusions; the "building blocks" of analysis and interpretation are evident)
- Findings/conclusions 'make sense'/have a coherent logic
- Findings/conclusions are resonant with other knowledge and experience (this might include peer or member review)
- Use of corroborating evidence to support or refine findings (i.e other data sources have been used to examine phenomena; other research evidence has been evaluated)

(Source: Spencer et al, 2003)

For each of the item in the checklist, the following scale-point will be used:

Yes – 1 point; No – 0 point; Partially – 0.5 point

The resulting total quality score for each study ranged between 0 (very poort) and 7 (very good). The quality score can be used an indicator of whether a study is highly reliable or not since the information is useful during the synthesis of evidence.

8. Data extraction strategy

After the final selection of papers, data extraction activity will be carried out on all papers that passed the screening process. During this stage, we will extract all important information that will help us analyse the evidence. The principle researcher is responsible to read the full paper and complete the data extraction form.

8.1 Required Data

Data are coded on a number of different variables. This includes information on the publication and other important attributes of a study. Table 14 shows an example of data extraction completed for a paper authored by Williams et al. (2006).

Table 14: Data Extraction Form - Completed for Williams et al., 2006

Data item	Value	Additional notes
Year	2006	
Author	L. Williams, L. Layman, J. Osborne, & N. Katira	
Title	Examining the Compatibility of Student Pair Programmers	
Reference type (journal/ conference paper/ thesis/ unpublished work)	Conference paper	
Journal/conference name	AGILE 2006 Conference	
Publisher	IEEE Computer Society	
Country of Study	USA	
Setting	University	
Aim of study	To determine whether instructors can proactively form compatible pairs based upon personality types, learning style, skill level, programming self-esteem, work ethic and time management preference	

Type of study	Experiment	
Who are the subjects involved?	Undergraduate and Graduate Students	Minorities identified are African American, Hispanic, and Alaskan/American Indian
Sample size & population	1350 students	Freshmen, advanced undergraduate and graduate students involved in this study
Describe the tasks	All students were required to complete a web-based peer evaluation survey on the contribution and compatibility of their partner after completing each of the paired assignments. For CS1, students had to complete four assignments during the semester; for SE, students had multiple 2 to 3-week programming assignments; for OO class, there were 3 class projects.	PairEval is the tool used for the online survey (At the start of a semester, students completed an online Myers-Briggs Type Indicator (MBTI) test and an online Felder-Silverman learning style test.
Duration of study	2 phase study (2002-2005)	Between Fall 2002 and Fall 2005
If subjects are students, what course(s) involved?	Introduction to programming (CS1) Software Engineering (SE) and Object oriented language & system (OO)	CS1 and SE are both for undergraduate students; OO is for graduate students
Does the subjects required to work in pairs? If so, do they change partner?	For CS1 and SE, students are mandated to work in pair. However, it is optional for OO course. They were assigned to different partner at the completion of each assignment.	
Did they only use PP students? Or did they also use solo students?	Both pair and solo students are used.	
How was pair students configured?	Students were paired if both have: -different personality types -different learning styles -similar perceived and actual skill, programming self-esteem and time management skill	
Hypotheses/ Research Question	Pairs are more compatible if students with [..] are grouped together: H1: [different personality types] H2: [different learning styles] H3: [similar perceived skills] H4: [similar actual skill levels] H5: [similar programming self-esteem] H6: [similar time management skills]	
List the independent variables (intervention)	Personality types, learning styles, perceived skills, actual skills level, programming self-esteem and time management skills	
List the dependant Variables (DV or outcomes)	Compatibility of students who pair programmed	
Any context variables defined?	Not reported	
How was the variable under study being measured?	Personality type is measured using the MBTI	

	Learning style is measure using the Felder-Silverman LS Actual skill level is based on midterm scores, GPA and SAT.	
What is the research design used?	Experiment. Detailed research design was not mentioned.	
PP Compatibility		
Compatibility factor addressed	Personality types, learning styles, perceived skills, actual skills level, programming self-esteem and time management skills	
What are results of PP compatibility?	Students will be highly compatible and successful if paired randomly, without necessarily considering personality type, skill level, self-esteem, work ethic or time management skills.	
PP effectiveness		
How was PP effectiveness measured?	N/A	
What are results of PP effectiveness?	N/A	
Quality of code		
If quality is a variable being studied, how was quality measured?	N/A	
What are the results?	N/A	
Statistical analysis		
Statistical methods used to analyse data	Spearman rank-order, chi-square test, logistic regression	
Effect Size available/ calculable? If yes, what are the candidate variables and its values?	No	
Overall findings	Result supports all hypotheses	
Study Quality		
Was the article referred?	Yes	
Were the aim(s) of study clearly stated?	Yes	
Were the study participants or observational units adequately described?	Yes	
Were the data collections carried out very well?	Partially	
Were the potential confounders adequately controlled for in the analysis?	No	
Were the approach to and formulation of the analysis was well conveyed?	Partially	
Were the findings credible?	Partially	
Total quality score	4.5/7	

9. Data extraction process

The data extraction process involves entering data using the data extraction form (see Section 6.1). Each data extraction form is stored electronically where the *Study Identifier* shall be used as the filename. The study identifier consists of the publication year concatenated

with the first author's last name. If more than one paper is written by the same author for a particular year, an alphabet will be added at the end of the file name.

In the event where findings reported in a study are ambiguous, the main author will be contacted for further clarification. Correspondence with author(s) is recorded under the *Additional Notes* column in the data extraction form. A review meeting will be held between the principal researcher and the supervisor(s) during the middle phase of data extraction. For validation purposes, a sample comprising 20% of the total number of primary studies will be selected randomly and had their data extracted by the principal researcher and her main supervisor. The extracted data will be later compared in a review meeting attended by review team members. Whenever the data extracted differed, such differences will be discussed until consensus is reached.

10. Synthesis of the extracted data:

This section outlines the strategy to synthesize evidence based on the data extracted from each primary studies included in the SR. We plan to perform the synthesis using a quantitative summary where information captured from each study is tabulated in a relevant table.

10.1 Question 1

The question is "What evidence is there regarding compatibility factors that affect pair compatibility and/or PP's effectiveness as a CS/SE pedagogical tool?"

This question will be addressed by tabulating the studies as shown in Table 15. The aim is to aggregate empirical studies investigated factor affecting pair effectiveness/productivity or pair compatibility.

Table 15: Summary on study related to compatibility factor (Quantitative)

Study ID	Author(s)/ Yeat	Type of study	Compatibility factor	Study Outcomes	Course/ Students involved	Sample size	Quality score
S1	Williams et al. (2006)	Formal Experiment	Personality types, learning style, skill levels, programming self esteem, work ethic, and time management skills	Compatibility can be achieved if students were paired randomly. Three factors contribute to compatibility: perceived skill, work ethic, and learning style. Students were compatible with partner they perceived of similar or higher skill level, similar work ethic, and different learning styles. Other variables are not significant contributors.	CS1 (Undergrad) SE (Undergrad) OO (Graduate)	1350 (two-phased study)	4.5
..	…						

Table 16 will be used to present a summary of qualitative studies. Finally, Table 17 will be used for the purpose to list the compatibility factors investigated in PP studies included in the SR. The aim is to identify which compatibility factors most commonly investigated in PP studies and also to examine the pattern of findings from those studies.

Table 16: Summary on study related to compatibility factor (Qualitative)

Study ID	Author(s)/ Yeat	Research Method	Compatibility factor	Study Outcomes	Course/ Students involved	Sample size	Quality score
S11	Cao & Ramesh (2004)	Exploratory	Skill level	Pairing combination is the most effective when the level of competency between the partners is about similar or not too different.	Undergraduate Programming Course	23	5.5
…	…						

Table 17: Compatibility factor found in PP studies

No	Compatibility Factor	Study(s)
1	Personality type	*(list of study identifier)*
2.	Skill levels	
3.	Gender	
4.	Ethnicity	
5.	Learning styles	
6.	Work ethic	
7	Time management ability	
..	…	

10.2 Question 2

The question is "Which pairing configurations are considered as most effective looking at the compatibility factors obtained from Question 1?

The answer to this research question depends on the findings in Question 1. In particular, we will examine each compatibility factor found in Question 1 and present the results as shown in Table 18.

Table 18: Summary on effective pairing configuration

Compatibility factor	Study(s)	Pairing configuration	Findings
Perceived skills	S1	Paired students with similar or higher skill level	Students preferred to pair with a partner they perceived of similar or higher skill level.
Learning style	S1	Pair students of sensor and intuitor learning style	Pairing a sensor and an intuitor lead to a very compatible pair.
Work ethic	S1	Paired students with similar work ethic	Students preferred to work with someone who has similar intention to success in the course.
…	…		

10.3 Question 3

The question is "How was PP's effectiveness measured in PP studies and how effective has PP been when used within higher education settings?"

The synthesis of evidence on the first part of this question concerned on outlining the method or approach to measure effectiveness of PP. Data from relevant studies will be presented as shown in Table 19. Data from qualitative studies will be tabulated as per Table 20.

Table 19: Studies investigated PP's effectiveness (Quantitative)

Study ID	Type of study	Measure of effectiveness	Outcome(s)	Pair Vs Solo	Sample size	Course/Task & Duration	Quality score
S3	Formal Exp.	Design knowledge diffusion and enforcement	Pair design can diffuse the knowledge and help enforce the design knowledge better than solo programming. However, the skills and individuals abilities could seriously affect the effectiveness of the practice.	Yes	132 (two exp.)	SE (2 hours 45 minute)	7
…	…						

Table 20: Studies measuring PP effectiveness (Qualitative)

Study ID	Research Design	Purpose of study	Method(s)	Outcome(s)	Sample size	Quality score
S14	Field study (exploratory)	To explore factors that may affect the success of PP	Participant observation, questionnaires, semi-structured interviews	Efficiency, enjoyment, and perception of learning all showed positive result and favour towards PP.	58	5
…	…					

In the second part of this question, we aimed to quantify how effective has PP been when used in academic settings. For this purpose, a meta-analysis shall be conducted to see whether there is any heterogeneity found between the studies. However, the feasibility of running this analysis depends upon the availability of calculating the studies' effect size. Effect size can be calculated using the appropriate statistical procedure such as standardized mean difference or a correlation coefficient. The effect size indices indicate the "measures of practical significance or meaningfulness" of the study findings (Dyba et al., 2006). We will illustrate the findings using Forest Plot or Funnel plot to visualize the patterns of data. The software tools to be used for this purpose will be either proprietary software such Comprehensive Meta-Analysis (CMA) or open source software (such as RevMan or MIX).

10.4 Question 4

The question is "How was quality measured in the PP studies that used software quality as a measure of effectiveness?"

This question concerned on extracting studies that used quality metrics to measure PP's effectiveness. For example, a study may used *Lines of Code*, *number of test cases passed/failed* or *Project scores/grades* to measure effectiveness of PP practice [33]. Quality metrics identified from PP studies will be tabulated as shown in Table 21. We would also provide the summary of findings from qualitative studies in a separate table using the same template as in Table 21.

Table 21: Quality Metrics used in Quantitative PP Studies

Study ID	Type of study	Quality measure(s)	Summary of findings	Course(s) involved	Pair Vs. Solo	Sample Size
S25	Case studies	LOC (without comments), Comment Ratio (CS) and Coupling factor (CF)	LOC for PP teams is slightly lower than non-PP teams. PP teams had slightly lower CR and CF.	Software Praktikum	Yes	24
…	…					

11. Schedule for Review

Detail description of each review task:

Formulate Research Question(s) & Protocol development (Duration: 2 months)

- Identify the rationale for review, strategy to search for primary studies, define the study selection criteria, quality assessment checklist, strategy to extract data and strategy to synthesis the studies.
- Conduct a pilot review.
- Write up/Update the protocol for systematic review of pair programming studies in higher education settings.

Conducting the review (Duration: 6 months)

Activities	**Duration**
Identify relevant literature	3 weeks
Selection of studies	4 weeks
Data extraction	9 weeks
Assessment of studies' quality (parallel with data extraction process)	-
Summarize evidence & interpretation of findings	8 weeks

Writing up SR report/results (Duration: 1.5 month)

Figure 0.2 illustrates the process of our Systematic Review conduct.

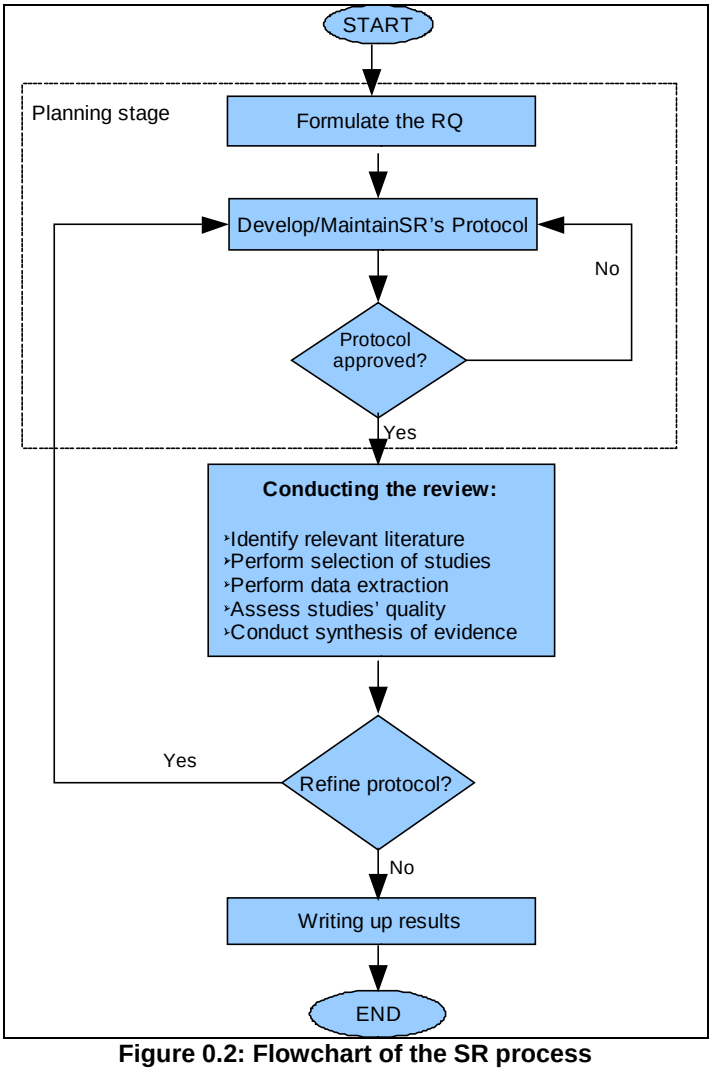


Figure 0.2: Flowchart of the SR process

Appendix A.2 Data Extraction Form

Date extracted: Study ID:

Data item	Value	Additional notes
Year		
Author		
Title		
Reference type (journal article/ conference paper/ thesis/ unpublished work)		
Journal/conference name		
Publisher		
Country of Study		
Setting		
Aim of study		
Type of study		
Who are the subjects involved?		
Sample size & population		
Describe the tasks		
Duration of study		
If subjects were students, which course(s) were they attending as part of the PP evaluation?		
Were the subjects required to work in pairs? If so, did they change partner?		
Did they only use PP subjects?		
How were the pairs allocated?		
Hypotheses/ Research Question(s)		
List the independent variable(s) (intervention)		
List the dependant variable(s) (DV or outcomes)		
Any context variable(s) defined?		
How was the dependant variable(s) under study being measured?		
What is the research design used?		
PP Compatibility		
Compatibility factor addressed		
What are results of PP compatibility?		
PP effectiveness		
How was PP's effectiveness measured?		
What are the results of PP effectiveness?		
Quality issues		
If quality is a variable being studied, how was it measured?		
Which were the results obtained when investigating quality?		
Statistical analysis		
Statistical method(s) used for data analysis		
Effect Size available/calculable? If yes, what are the candidate		

variables and their values?		
State the overall findings from the paper?		
Study Quality (Answer either: Yes/No/Partially) Yes=1; No = 0; Partially = 0.5		
Was the article referred?		
Were the aim(s) of the study clearly stated?		
Were the study participants or observational units adequately described? For example, students' programming experience, year of study etc.		
Were the data collection carried out very well? For example, discussion of procedures used for data collection, and how the study setting may have influenced the data collected		
Were the potential confounders adequately controlled for in the analysis?		
Were the approach to and formulation of the analysis was well conveyed? For example, description of the form of the original data, rationale for choice of method/tool/package.		
Were the findings credible? For example, the study was methodologically explained so that we can trust the findings; findings/conclusions are resonant with other knowledge and experience.		
Total quality score:		

Additional Comments:

Appendix A.3 List of Included Studies

[S1] Al-Kilidar, H., Parkin, P., Aurum, A., & Jeffery, R. (2005). Evaluation of effects of pair work on quality of designs. *Australian Software Engineering Conf.*, 78-87.

[S2] Balijepally, V. (2006). *Task complexity and effectiveness of pair programming: An experimental study.* Unpublished Ph.D., The University of Texas at Arlington, United States -- Texas.

[S3] Bellini, E., Canfora, G., Garcia, F., Piattini, M., & Visaggio, C. A. (2005). Pair designing as practice for enforcing and diffusing design knowledge. *Journal of Software Maintenance and Evolution-Research and Practice, 17*(6), 401-423.

[S4] Berenson, S. B., Slaten, K. M., Williams, L., & Ho, C.-w. (2004). Voices of women in a software engineering course: Reflections on collaboration. *Journal of Educational Resources in Computing (JERIC), 4*(1).

[S5] Bipp, T., Lepper, A., & Schmedding, D. (2008). Pair Programming in Software Development Teams An Empirical Study of its Benefits. *Information and Software Technology, 50*(3), 231-240.

[S6] Canfora, G., Cimitile, A., & Visaggio, C. A. (2004). *Working in pairs as a means for design knowledge building: An Empirical Study.* Paper presented at the 12th IEEE International Workshop on Program Comprehension (IWPC'04).

[S7] Canfora, G., Cimitile, A., & Visaggio, C. A. (2005). Empirical study on the productivity of the pair programming. *Extreme Programming and Agile Processes in Software Engin. Proc. 6th Int'l Conf., XP 2005, LNCS. 3556, Springer-Verlag*, 92-99.

[S8] Canfora, G., Cimitile, A., Garcia, F., Piattini, M., & Aaron VIsaggio, C. (2005). Confirming the influence of educational background in pair-design knowledge through experiments. *ACM Symposium on Applied Computing*, 1478 - 1484

[S9] Canfora, G., Cimitile, A., Garcia, F., Piattini, M., & Visaggio, C. A. (2006). Performances of pair designing on software evolution: a controlled experiment. *10th European Conference on Software Maintenance and Reengineering,* 195-202

[S10] Srikanth, H., Williams, L., Wiebe, E., Miller, C., & Balik, S. (2004). On Pair Rotation in the Computer Science Course. *17th Conference on Software Engineering Education and Training (CSEET'04)*, 144 - 149.

[S11] Lan, C., & Ramesh, B. (2004). An exploratory study on the effects of pair programming. *8th Int'l Conf. on Empirical Assessment in Software Engineering (EASE 2004) Workshop - 26th International Conference on Software Engineering.* , 21-28.

[S12 Carver, J. C., Henderson, L., He, L., Hodges, J., & Reese, D. (2007). Increased Retention of Early Computer Science and Software Engineering Students Using Pair Programming. *20th Conf. on Software Engin. Education & Training, (CSEET '07)*, 115-122.

[S13] Chao, J., & Atli, G. (2006). Critical Personality Traits in Successful Pair Programming. *AGILE'06, IEEE Computer Society*, 89-93.

[S14] Chaparro, E. A. (2005). Factors affecting the perceived effectiveness of pair programming in higher education. *17th Workshop of the Psychology of Programming Interest Group, Sussex University*, 5-18.

[S15] Cliburn, D. C. (2003). Experiences with Pair programming at a Small College. *Journal of Computing Sciences in Colleges, 19*(1), 20-29.

[S16] DeClue, T. H. (2003). Pair Programming and Pair trading: Effects on Learning and motivation in a CS2 courses. *Journal of Computing Sciences in Colleges, 18*(5), 49-56.

[S17] Domino, M. A., Collins, R. W., R., H. A., & Cohen, C. F. (2003). Conflict in Collaborative Software Development. *Proceedings of the 2003 SIGMIS Conference on Computer Personnel Research: Frreedom in Philadelphia-leveraging differences and diversity in the IT workforce SIGMIS CPR'03*, 44-51.

[S18] Domino, M. A., Webb Collins, R., & R. Hevner, A. (2007). Controlled experimentation on adaptations of pair programming. *Information Technology and Management, 8*(4), 297-312.

[S19] Freeman, S. F., Jaeger, B. K., & Brougham, J. C. (2003). Pair programming: More learning and less anxiety in a first programming course. *ASEE Annual Conference Proceedings*, 8885-8893.

[S20] Gehringer, E. F. (2003). A Pair-Programming experiment in a Non-Programming

courses. *OOPSLA'03*, 187 - 190.
[S21] Hanks, B., McDowell, C., Draper, D., & Krnjajic, M. (2004). Program quality with pair programming in CS1. *SIGCSE Bulletin, 36*(3), 176-180.
[S22] Hanks, B. (2006). Student attitudes toward pair programming. *Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE06)*, 113-117.
[S23] Heiberg, S., Puus, U., Salumaa, P., & Seeba, A. (2003). Pair-Programming Effect on Developers Productivity. *Extreme Programming and Agile Processes in Software Engineering - Proc. 4th International Conference, XP 2003, Springer-Verlag, LNCS 2675*, 215-224.
[S24] Ho, C.-w. (2004). *Examining Impact of Pair Programming on Female Students* (No. TR-2004-20). Raleigh, NC: North Carolina State University.
[S25] Ciolkowski, M., & Schlemmer, M. (2002). Experiencing with a Case Study on Pair Programming. *in Proceedings of the Workshop on Empirical Studies in Software Engineering (PROFES 2002), Finland*.
[S26] James, S. D., & Hansen, J. C. (2002). Student-based pair programming: an examination. *Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, 8*, 485-489.
[S27] Janes, A., Russo, B., Zuliani, P., & Succi, G. (2003). An empirical analysis on the discontinuous use of pair programming. *Extreme Programming and Agile Processes in Software Engineering - Proc. 4th International Conference, XP 2003. LNCS 2675, Springer-Verlag*, 205-214.
[S28] Katira, N., Williams, L., Wiebe, E., Miller, C., Balik, S., & Gehringer, E. (2004). On understanding compatibility of student pair programmers. *SIGCSE Bulletin, 36*(1), 7-11.
[S29] Katira, N., Williams, L., & Osborne, J. (2005). Towards Increasing the Compatibility of Student Pair Programmers. *27th International Conference on Software Engineering (ICSE'05)*, 625-626.
[S30] Kuppuswami, S., & Vivekanandan, K. (2004). The effects of pair programming on learning efficiency in short programming assignments. *Informatics in Education, 3*(2), 251-266.
[S31] Layman, L., Williams, L., Osborne, J., Berenson, S., Slaten, K., & Vouk, M. (2005). How and Why Collaborative Software Development Impacts the Software Engineering Course. *Proceedings of the 35th Annual Conference Frontiers in Education (FIE '05)*, T4C-9-T4C-14.
[S32] Layman, L. (2006). Changing students' perceptions: an analysis of the supplementary benefits of collaborative software development. *Proceedings 19th Conf. on Software Engineering Education & Training, IEEE Computer Society*, 159 - 166
[S33] Lui, K. M., & Chan, K. C. C. (2006). Pair programming productivity: Novice-novice vs. expert-expert. *Int'l Journal of Human-Computer Studies, 64*(9), 915-925.
[S34] Madeyski, L. (2005). Preliminary analysis of the effects of pair programming and test-driven development on the external code quality. In K. Zieliński & T. Szmuc. (Eds.), *Software Engineering: Evolution and Emerging Technologies* (pp. 113–123): IOS Press.
[S35] Madeyski, L. (2006). The impact of pair programming and test-driven development on package dependencies in object-oriented design - an experiment. *in Proceedings 7th International Conference Product-Focused Software Process Improvement (PROFES 2006), LNCS 4034, Springer-Verlag*, 278-289.
[S36] Madeyski, L. (2007). On the Effects of Pair Programming on Thoroughness and Fault-Finding Effectiveness of Unit Tests. In Jurgen Munch & P. Abrahamsson (Eds.), *Product-Focused Software Process Improvement* (Vol. 4589/2007, pp. 207-221). Berlin: Springer-Verlag.
[S37] McDowell, C., Werner, L., Bullock, H., & Fernald, J. (2002). The effects of pair-programming on performance in an introductory programming course. *ACM. SIGCSE Bulletin, 34*(1), 38-42.
[S38] McDowell, C., Hanks, B., & Werner, L. (2003). Experimenting with pair programming in the classroom. *ACM. SIGCSE Bulletin, 35*(3), 60-64.
[S39] McDowell, C., Werner, L., Bullock, H. E., & Fernald, J. (2003). The Impact of Pair Programming on Student Performance, Perception and Persistence. *Proceedings of the 25th International Conference on Software Engineering*

(ICSE'03), 602-607.
[S40] Mendes, E., Al-Fakhri, L. B., & Luxton-Reilly, A. (2005). Investigating pair-programming in a 2nd-year software development and design computer science course. *SIGCSE Bulletin, 37*(3), 296-300.
[S41] Mendes, E., Al-Fakhri, L., & Luxton-Reilly, A. (2006). A replicated experiment of pair-programming in a 2nd-year software development and design computer science course. *Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE06)*, 108-112.
[S42] Muller, M. M., & Padberg, F. (2004). An empirical study about the feelgood factor in pair programming. *Proceedings of the 10th International Symposium on Software Metrics,(METRIC 2004)*, 151-158.
[S43] Muller, M. M. (2004). Are reviews an alternative to pair programming? *Empirical Software Engineering, 9*(4), 335-351.
[S44] Muller, M. M. (2005). Two controlled experiments concerning the comparison of pair programming to peer review. *Journal of Systems and Software, 78*(2), 166-179.
[S45] Muller, M. M. (2006). Do programmer pairs make different mistakes than solo programmers? *Journal of Systems and Software, 80*(9), 1460-1471.
[S46] Muller, M. M. (2006). A preliminary study on the impact of a pair design phase on pair programming and solo programming. *Information and Software Technology, 48*(5), 335-344.
[S47] Nawrocki, J., & Wojciechowski, A. (2001). Experimental Evaluation of Pair Programming. *Proceedings European Software Control and Metrics (Escom)*, 269-276.
[S48] Ahren, T. C. (2005). *Work in progress - effect of instructional design and pair programming on student performance in an introductory programming course.* Paper presented at the 35th Annual Conf. on Frontiers in Education (FIE '05).
[S49] Phongpaibul, M., & Boehm, B. (2006). An Empirical Comparison Between Pair Development and Software Inspection in Thailand. *Proceedings of the 5th ACM-IEEE Int'l Symp. on Empirical Software Engineering (ISESE'06)*, 85 - 94
[S50] Sfetsos, P., Stamelos, I., Angelis, L., & Deligiannis, I. (2006). Investigating the impact of personality types on communication and collaboration-viability in pair programming - an empirical study. *Extreme Programming and Agile Processes in Software Engin.- Proc. 7th Int'l Conf. XP 2006, LNCS 4044,* 43-52.
[S51] Slaten, K. M., Droujkova, M., Berenson, S. B., Williams, L., & Layman, L. (2005). Undergraduate student perceptions of pair programming and agile software methodologies: verifying a model of social interaction. *Proceedings of the Agile 2005, IEEE Comput. Soc.*, 323-330.
[S52] Xu, S. & Rajlich, V. (2005). *Pair Programming in Graduate Software Engineering Course Projects*. Paper presented at the 35th Annual Conference on Frontiers in Education (FIE '05), F1G-7-F1G-12.
[S53] Shaochun, X., & Rajlich, V. (2006). Empirical validation of test-driven pair programming in game development. *Proceedings of the 5th IEEE/ACIS Int'l Conf. on Computer Information Science - In Conjunction with 1st IEEE/ ACIS Workshop on Component-Based Software Engineer Architecture and Reuse,* 500-505.
[S54] Thomas, L., Ratcliffe, M., & Robertson, A. (2003). Code Warriors and Code-a-Phobes: A Study in Attitude and Pair Programming. *SIGCSE Bulletin, 35*(1), 363-367.
[S55] Tomayko, J. E. (2002). A Comparison of Pair Programming to Inspections for Software Defect Reduction. *Computer Science Education, 12*(3), 213-222.
[S56] VanDeGrift, T. (2004). Coupling pair programming and writing: learning about students' perceptions and processes. *ACM SIGCSE Bulletin, 36*(1), 2-6.
[S57] Vanhanen, J., & Lassenius, C. (2005). Effects of pair programming at the development team level: an experiment. *Proceedings of the 2005 International Symposium on Empirical Software Engineering (ISESE 05), IEEE CS Press*, 336 - 345.
[S58] Van Toll Iii, T., Lee, R., & Ahlswede, T. (2007). Evaluating the Usefulness of Pair Programming in a Classroom Setting. *Proceedings of the 6th IEEE/ACIS Int'l Conf. on Computer and Information Science (ICIS 2007)*, 302-308.
[S59] Williams, L. A., & Kessler, R. R. (2000). The effects of "pair-pressure" and "pair-learning" on software engineering education. *13th Conference on Software Engineering Education and Training, IEEE Comput. Soc.*, 59-65.
[S60] Williams, L., Kessler, R. R., Cunningham, W., & Jeffries, R. (2000). Strengthening the

case for pair programming. *IEEE Software, 17*(4), 19-25.
[S61] Williams, L., Wiebe, E., Yang, K., Ferzli, M., & Miller, C. (2002). In Support of Pair Programming in the Introductory Computer Science Course. *Computer Science Education, 12*(3), 197-212.
[S62] Williams, L., McDowell, C., Nagappan, N., Fernald, J., & Werner, L. (2003). Building pair programming knowledge through a family of experiments. *Proceedings 2003 Int'l Symposium on Empirical Software Engineering (ISESE 2003)*, 143-152.
[S63] Williams, L., Layman, L., Osborne, J., & Katira, N. (2006). Examining the Compatibility of Student Pair Programmers. *Proceedings of the Conference on AGILE 2006 (AGILE'06), IEEE Computer Society*, 411-420.
[S64] Winkler, D., & Biffl, S. (2006). An empirical study on design quality improvement from best-practice inspection and pair programming. In J. Munch & M. Vierimaa (Eds.), *Product-Focused Software Process Improvement* (Vol. 4034, pp. 319-333). Berlin: Springer-Verlag Berlin.
[S65] Nawrocki, J. R., Jasinski, M., Olek, L., & Lange, B. (2005). Pair programming vs. side-by-side programming. *Proceedings 12th European Conference on Software Process Improvement (EuroSPI 2005), LNCS 3792, Springer-Verlag*, 28-38.
[S66] Howard, E. V. (2006). Attitudes on using pair-programming. *Journal of Educational Technology Systems, 35*(1), 89-103.
[S67] Mujeeb-u-Rehman, M., Xiaohu, Y., Jinxiang, D., & Abdul Ghafoor, M. (2005). Heterogeneous and homogenous pairs in pair programming: an empirical analysis. *Proceedings of the 2005 Canadian Conference on Electrical and Computer Engineering (CCECE/CCGEI)*, 1116-1119.
[S68] Madeyski, L. (2006). Is external code quality correlated with programming experience or feelgood factor? *Proceeding of the7th International Conference on Extreme Programming and Agile Processes in Software Engineering (XP 2006), LNCS 4044, Springer-Verlag*, 65-74.
[S69] Simon, B., & Hanks, B. (2007). First year students' impressions of pair programming in CS1. *Proceedings of the 3rd International Workshop on Computing Education Research (ICER'07)*, 73-86.
[S70] Williams, L., Layman, L., Slaten, K. M., Berenson, S. B., & Seaman, C. (2007). On the Impact of a Collaborative Pedagogy on African American Millennial Students in Software Engineering. *Proceedings of the 29th International Conference on Software Engineering (ICSE 2007)*, 677-687.
[S71] Hanks, B. (2007). Problems encountered by novice pair programmers. *Proceedings of the 3rd Int'l Workshop on Computing Education Research (ICER'07)*, 159-164.
[S72] Williams, A. T. (2007). *Pair formation in CS1: Self-selection vs Random pairing.* Unpublished Ph.D. Dissertation, Pace University, New York, USA.
[S73] Choi, K. S. (2004). *A Discovery and Analysis of Influencing Factors of Pair Programming.* Unpublished Ph.D. Dissertation, New Jersey Institute of Technology, USA.
[S74] Gevaert, H. (2007). *Pair programming unearthed.* Unpublished M.Sc. thesis, University of Manitoba, Canada.

Appendix A.4 Summary of PP studies

Table A.4.1: RQ1 – Summary on study related to compatibility factor (Quantitative)

ID	Author(s)	Type of study	Compatibility factor	Outcomes	Course(s)/ students involved	Sample size	Task & duration	Quality score
S8	Canfora et al. (2005)	Formal Exp.	Skill level	Results from the first experiment showed that forming pairs with same educational background emphasizes the expected benefits of pair designing. Coupling two different academic backgrounds does not seem to improve the performance. In the replication study, results showed that the background of pair's components affects the knowledge built. Pair designers from scientific background seem to build greater knowledge. Thus, educational background is determinant in increasing the knowledge built by practice	SE (G)	14 pairs 16 solo (in first exp.) 32 pairs, 32 solo (in replicated exp.)	Subjects were given maintenance tasks to improve the UML design of the system Over 2 hours 45 min spent for the first experiment and 2 hours for the replicated experiments.	5.5
S13	Chao & Atli (2006)	Survey & Formal Exp.	Personality traits	Pair Programming (PP) success (code quality and pair compatibility) is not influenced by differences in personality traits.	CS1 (UG)	29 pairs	One week to complete programming assignment.	5
S15	Cliburn (2003)	Case Studies	Skill level	PP works best when the skill levels of partners are similar.	CS1 (UG)	27 students	5 programming projects and weekly assignment.	4.5
S22	Hanks (2006)	Survey	Confidence level	The most confident students like PP the most; the least confident students like it the least, thus contradicts the findings of Thomas (2003). The result also suggested that instructor may have a significant effect on students' attitude toward PP.	CS1 (UG)	134 students	Programming assignments; not mentioned about the duration	4.5
S23	Heiberg et al. (2003)	Formal Exp.& Exploratory	Personality type	The individual personality traits do not have significant consequences to pair programming performance.	OO (UG)	110 students	2 groups of pair and solo to solve part of a games application; 2 sessions, once a week.	6
S28	Katira et al. (2004)	Formal Exp.	Personality type, actual skill level, perceived technical competence, and self-esteem	Students were compatible if being paired randomly. Students have a preference to pair with a student they perceive to be of similar technical competence. Freshmen prefer to work with partners of different MBTI skills. Graduate students prefer to collaborate with partners of similar actual skill level. Students' self esteem has no correlation with pair compatibility.	CS1 (UG) SE (UG) OO (G)	564	To complete assignments in a closed lab (CS1); to complete 5 assignments (SE); to work on 3 class projects with no associated lab section (OO)	5.5

S29	Katira et al. (2005)	Formal Exp.	Personality types, actual and perceived skills, self-esteem, gender and ethnicity	Compatibility was significantly influenced by the perceived skill and actual skill. Personality types and self-esteem are not critical for pair compatibility. Compatibility can be achieved if the pair consists of: 1) similar perceived skills level, 2) similar actual skills level, 3) female students, 4) minority students	SE (UG) OO (G)	361	Students were given multiple 2-3 week assignments over 3 academic semesters	4.5
S32	Layman (2006)	Survey	Personality type, learning style, work ethic, time management ability	Students who dislike pairing were those who experienced having incompatible partner. Students who possess sensing-intuitive learning style showed higher preference to work in pairs similar to those extraverts of MBTI skill. Students in the group who disliked collaborating were reflective learners, introverts and strong coders. Students' preference whether to pair or not was highly affected by the compatibility of the pair. Personality and learning style had little influence towards perception of collaboration.	SE (UG)	119	Had to solve 2 solo assignment, one paired assignment, and 6-week group project in the first semester. Second sem. involving 2 paired assignments, one solo assignment, and 6-week group project	5
S42	Muller & Padberg (2004)	Formal Exp.	Skill level and "Feelgood" factor	Programming experience has no correlation with pair performance. Pair performance is correlated with the "feelgood" factor of the pair. However, the study cannot determine whether the performance is originally due to the "feelgood" factor or because the developers have the impression that they are performing well.	XP (UG)	19 pairs	Had to solve programming tasks: Polynomial and Shuffle puzzle; study involved 2 acad. semesters;	5.5
S50	Sfetsos et al. (2006)	Formal Exp.	Personality type and temperament type	Pairs of mixed-personalities and temperaments showed better performance and collaboration-viability. They achieve better points on assignments and shorter time to complete the tasks.	SE (UG)	84	To design, code and test two tasks on Coffee Machine Design using Java within two and a half hours	6.5
S54	Thomas et al. (2003)	Formal Exp.	Confidence level	Students whom identified themselves as "warrior" prefer to work alone and less enjoyed pair programming. Paired students of similar confidence level can cause greater performance.	CS1 (UG)	60 approx.	Had to solve simple coding problem during the lab hour	5.5
S63	Williams et al. (2006)	Formal Exp.	Personality types, learning styles, skill levels, programming self-esteem, work ethic and time management	Compatibility can be achieved if students were paired randomly. Paired students were compatible with partner they perceived of similar or higher skill level, and having similar work ethic. Overall results on personality and learning style showed partial support in predicting compatibility. However, pairing of sensor and intuitor produce a very compatible pair. Other variables such as actual skill level, self-esteem and time management skills are not significant	CS1 (UG) SE (UG) OO (G)	1350	Two-phased study; CS1 students had to complete 4 assignments during the semester; SE students had multiple 2 to 3-week programming assignment; OO students had 3 class projects but no lab sections. Students	4.5

			skills	contributors.			in CS1 and SE must work in a closed lab to do their assignments.	
S73	Choi K. S.	Formal Exp.	Personality, communication skills, and gender	Findings showed that personality differences do not give significant impact on communication, satisfaction, confidence, and compatibility level of the students. Gender differences were also found not significant on the pairs communication, satisfaction, confidence, and compatibility level.	Unk.	128	4 programming problems within 90 minutes.	5.5
S74	Gevaert H.	Formal Exp	Personality	Results showed that personality does not significantly affect the efficiency of students who paired. However, based on neuroticism domain, the higher the level of neuroticism, the lower the effectiveness of solo. From the extraversion dimension, there was no correlation with their performance at PP and SP. Pair satisfaction was found related to the personality than to the person's skills.	Unk.	28	Two programming problems to be solved in 30 minutes for the first task, and 60 minutes for the second task.	5
UG - undergraduate student **G** - **graduate students**			**SE** – Software Engineering **OO** – Object Oriented Programming **CS1** - Introduction to Programming					

Table A.4.2 RQ 1 – Summary on study related to compatibility factors (Qualitative)

ID	Author(s)	Type of study	Compatibility factor	Outcomes on pair compatibility	Course(s)	Sample size	Task & duration	Quality score
S11	Cao & Ramesh (2004)	Exploratory	Skill level	Pairing combination is the most effective when the level of competency between the partners is about similar or not too different.	Undergraduate Programming Course	23	9 weeks to develop an online Univ. registration system.	5.5
S14	Chaparro et al. (2005)	Field study/ Exploratory	Skill level, type of role, type of tasks	Skill level and the type of tasks are the most influential factors affecting the perceived effectiveness of pair programming. The study suggested that the skill level gap between the partners should not be too broad.	OOP (Postgraduate)	58	8 lab sessions 4 paired and 4 solo sessions involving small programming exercises.	5
S58	Van Toll, T. et al. (2007)	Case studies	Skill level	The findings confirmed Jensen's theory that PP works best when the programmers are of slightly different skill level (the skill level gap should not be too broad). PP is an effective tool to increase both satisfaction and learning ability.	N/A	N/A	Four programming projects; not mention about duration	3.5

Table A.4.3 RQ 2 – PP Studies measuring PP's effectiveness (Quantitative studies)

ID	Type of study	Measure of effectiveness	Outcome(s)	Pair Vs Solo?	Sample size	Course(s)	Task & duration	Quality score
S1	Formal Exp.	Using quality metric defined in ISO/IEC 9126.	Pair programmers produce significantly better design in terms of functionality, usability and portability. However, for more complex requirements, there was no significant difference in any quality characteristics between paired and solo designers.	Yes	150	Unk.	A 6-week project to design "planning" and "tracking" module of a web based network project management tool. Each project to be completed in 3 weeks.	6
S2	Formal Exp.	Satisfaction, confidence	Paired students produce higher quality software. Task complexity does not affect the quality of software produced by either pair or solo programmers. The satisfaction and confidence was also higher for pairs compared to the individuals.	Yes	120	Information System	Develop Java application within 3 hours	7
S3	Formal Exp.	Design knowledge diffusion and enforcement	Pair design can diffuse the knowledge and help enforce the design knowledge better than solo programming. However, the skills and individuals abilities could seriously affect the effectiveness of the practice.	Yes	132 (two exp.)	SE	Two maintenance tasks in 2 hours 45 minutes in order to reduce complexity and improve readability of UML design	7
S5	Formal Exp.	Task performance (LOC and quality of software produced) and acceptance of PP	In terms of code complexity, there was a trend that paired teams produce less complex programs. According to expert judgement, the code of the paired teams was rated a little bit better. Its readability and understandability were somewhat higher.	Yes	95	Software Praktikum	The 1st study involved a card game and management of a cocktail bar. The 2nd study involved two projects: a quiz game and the simulation of the elevators in a multi-story building. Paired and solo team in both studies solves the same tasks.	6
S6	Formal Exp.	Level of knowledge building	Level of knowledge building is higher for pairs than solos.	Yes	45	SE	3 runs of experiments; to deliver design document, including use case diagram, class diagrams and interaction diagrams, one for each run	5
S7	Formal Exp.	Time spent programming	The results showed that the same developer decreases the time for developing a task when moves to pair programming from solo programming.	Yes	24	MUTS	Two runs of experiments where students were required to develop two applications, one for each run.	5

S8	Formal Exp.	System's knowledge	Pair designers from scientific background seem to build greater knowledge. Thus, educational background is determinant in increasing the knowledge built by practice	Yes	64 (first exp.) 96 (second exp.)	SE	Maintenance tasks to improve the UML design of the system Over 2 hours 45 min spent for the first experiment and 2 hours for the replicated experiments.	5.5
S9	Formal Exp.	The effort spent, and the achieved quality	Pair designing is able to decrease the time spent to accomplish task. The quality achieved with pair designing is significantly greater than the solo designing.	Yes	70	Unk.	2-hour evolution tasks on the design of a software system. There were 2 assignments; each consists of two evolution tasks.	6
S12	Formal Exp.	Retention rate	Students who pair programmed showed an increase to remain in the CS/SE/CE majors (Pairs/ solo was done in separate semester)	Yes	104	Intro. to Computer Prog.	Students are required to participate in a 3-hour lab session each week, to solve programming assignments.	6.5
S13	Surney & formal Exp.	Quality of code (rated by experts)	Quality of code does not determine by the personality of pair developers.	All paired	118	CS1	One week to complete the programming assignment	5
S15	Case Studies	Learning effect (Final course grade, enjoyment)	PP produces better projects in less time, and results in reduce workload for teaching staffs. Student evaluation PP seems to be very positive; most students enjoy the class when paired (76.9%). The results suggest that PP is an effective tool for introductory programming course and that PP works best when partners are at the same ability level.	Yes	17	CS1	Students were given 5 programming projects besides weekly assignment.	4.5
S17	Formal Exp.	Number of test cases failed, coding error	Cognitive ability of developers and faithfulness to the method is significant to determine the performance of pair programmers. Besides, PP requires effective conflict management skills.	All paired	14	Unk.	Three programming exercises in a 4-week time frame	5
S18	Formal Exp.	Based on tasks performance and user satisfaction	PP works best when working face-to-face instead of virtually. Developers showed higher level of satisfaction when pair programming in face-to-face work setting compared to working alone or in virtual setting.	Yes	216	Unk.	Students were assigned two experimental tasks, Task I and Task II (Task II is more difficult). Subjects were asked to write pseudocode using their own knowledge in specific programming language. 45 minutes were allotted for the completion of each tasks.	7

S19	Survey	Academic performance (quiz, final exam and course grade), time spent on programming project	There was no significant difference between pair and solo students in terms of performance on the quizzes, final exam, and overall course grade. The time spent on project shows no significant difference between pair and solo students.	Yes	128	Engin. Problem Solving with Computation	Students worked solo for the 1st coding assignment but worked in pair for the 2nd assignment. They can chose whether to pair or worked solo in the 3rd assignments. Programming assignments given on weekly basis.	4.5
S20	Formal Exp.	Project scores, and students' perception	There is no significant difference between performance of pair and solo students in their course projects. Students' perception toward PP was very good. (they ranked it 3.17 of 4 scale)	Yes	101	Computer Design & Technology	To develop three simulation projects in one academic semester	4.5
S21	Formal Exp.	Program quality	Pair students produce programs with greater functionality than the solo students. They were also more confident and more satisfied with the programming process. There was not enough evidence to confirm that pair students write programs with better design, and show better understanding of programming concepts.	Yes	162 approx.	CS1	5 assignments in one semester but only the last three were evaluated. Program 3: a card game blackjack; Program 4: simple dice game using a one-dimensional array. Program 5: a text-based version of Minesweeper game.	5
S22	Survey	Confidence, enjoyment	Students responded with positive attitude towards PP. They believe that PP helps them learn more and make the class more fun.	Only paired	115	Intro. Prog.	Programming assignments	4
S23	Formal Exp. Exploratory	Technical productivity	There was no difference in productivity between pair and non-pair programmers. Productivity was measured based on number of solutions that satisfy test cases, and average percentage of satisfied test cases per pair.	Yes	75	OOP	Students divided into two groups of pair and solo to solve a component of a game environment; two sessions, once a week.	6
S25	Case Study.	Code Quality, effort used, and students' subjective impression	Code quality measured using LOC, Comment ratio (CR), and Coupling factor (CF). Effort was measured based on effort per person in hours. Results: Teams using PP had slightly lower LOC, lower CR, and lower CF compared to non-PP teams. Effort for PP teams is slightly higher than non-PP teams. The study was not able to find strong support that PP helps increase quality, and that PP team spent more effort. Students have good impression about using PP.	Yes	24	Software praktikum	Implement changes for existing software written in Java known as "web based quiz" system over 13 weeks	5
S26	Formal Exp.	Knowledge transfer	81% of the students indicated that PP facilitates their learning.	All paired	34	Data Comm. & Networking	Course project and assignments for the duration of half of the semester	4

S27	Formal Exp.	Knowledge and skills transfer	80% of students expressed the benefits of PP in transfer of knowledge and skills during the internship.	All paired	15	Summer internship	Students were required to do project during summer internship program	6
S30	Formal Exp.	Learning efficiency (design score, time spent and test score)	Pair students produce quality design better than solo students. They were also able to complete the lab exercises in a shorter duration. Knowledge and programming skills were also higher for paired students.	Yes	58 pairs, 98 solo	Prog. languages, Internet & Database prog.	Lab exercises and a written test. Lab exercises lasted in 3 hours duration without a break.	7
S31	Survey	Perception that pairing saves time and, increased code quality.	Students with lower self-confidence in their programming skills prefer to work with other studentsMost students but higher confidence students perceived that pairing helps in reducing errors, thus increased the code quality.	Yes	119	SE	Assignments emphasize on understanding requirements, creating a design, and writing code and tests. Data collected in two academic semesters.	4.5
S32	Survey	Attitude towards collaboration	Students overwhelmingly showed positive change in their preference to collaborate (i.e after experiencing PP, students prefer to work with another student) Students also feel more organized and able to complete the assignment faster with pair. Those who disagree with pair were having incompatible partner and scheduling conflict.	All paired	78	SE	2 individual assignments, one paired assignment and 6-week group project during the 1st semester. The 2nd semester involved 2 paired assignments, 1 solo assignment, and 6-week group project. Assignments in both semesters lasted about 1-2 weeks	5
S33	Formal Exp.	PP productivity (time spent and software quality	PP Productivity diminishes when pairs keep solving the same problem. The novice-novice pairs against novice solos are much more productive in terms of elapsed time and software quality than expert-expert pairs against expert solos. PP effectively helps developers solve unfamiliar programming problems.	Yes	40	Agile Software Development and XP	Students were asked to write a FIFO warehouse using SQL server and ASP within an academic semester.	7
S34	Formal Exp.	The external code quality (measured by a number of acceptance tests passed/ NATP)	Code quality was significantly affected by software development approach (i.e. Classical/ test driven development on solo/PP). Classical approach (pair & solo) achieved better results compared to a test-driven approach. Test-driven development decreases the external code quality when used by either pair or solo students. There was no difference in the external code quality when PP was used instead of solo programming.	Yes	188	Prog. in Java (PIJ)	Eight lab sessions (90 minutes per each) to develop finance-accounting system	7
S35	Formal Exp.	Quality of OO design (Martin's package level dependency metrics)	Results indicate imperfect package design regardless of software development method used. The study found that package level design quality indicators were not significantly affected by development method (pair or solo).	Yes	122	Prog. in Java (PIJ)	Eight lab sessions (90 minutes per each) to develop finance-accounting system	7

S36	Formal Exp.	Thoroughness in code coverage & the test cases quality	The results do not support the positive impact of PP on testing to make it more effective and thorough.	Yes	63	Prog. in Java (PIJ)	Eight lab sessions (90 minutes per each) to develop finance-accounting system	7
S37	Formal Exp.	Academic performance (assignment and final exam scores), course completion rate	Pair students score significantly higher marks in programming assignment but score slightly lower than solo students in the final exam. Course completion rate was significantly higher in the pairing class.	Yes	600 approx.	CS1	Five comparable programming assignments given to students in pairing section (Fall 2000), and non-pairing section (Spring 2001).	6
S38	Formal Exp.	Final exam performance, programming assignments scores, program quality and the total time spent on programming	No significant difference in exam scores between pairing and non-pairing students.The scores on programming assignments were significantly higher for pairing sections than the solo sections. Program quality produced by pair students is holistically better compared with solo students but slightly lower in terms of functionality and style. Overall results showed paired students produced a significantly better quality program than the solo students. In terms of time spent programming, there was no significant difference between the two groups.	Yes	216 95	Advanced prog., abstract data types	A series of experiments were conducted over three academic semesters with two of the studies on voluntarily basis and no monitoring done on the PP practice. Students were given option whether to pair or work solo for the assignments.	4.5
S39	Formal Exp.	Completion and pass rates, course performance (programming assignment score and final exam score)	Paired students were significantly more likely to complete the course (90% Vs 80.4%) and more likely to pass the course compared to solo students (72.3% Vs 62.8%). Average programming scores for pair students were higher than solo students. However, the final exam score does not affected by whether the students paired or not.	Yes	555	CS1	Students worked in pair for all assignments over 2 semesters (fall and winter) and students enrolled in Spring worked solo. Paired students required to submit 5 programming assignments and a log entry indicating the amount of time spent on the assignment, their level of confidence, enjoyment and satisfaction with the process.	5
S40	Formal Exp.	Students' success rate and students enjoyment	Results showed that success rate (students who passed with grade 'C' and above) is greater for paired students. Paired students also received better score in all their assignments, test and final exam. Majority students (74%) enjoyed the PP experience.	Yes	300	Software Design and Construction	A 12-week course; students had to deliver 3 assignments: draw UML diagrams, convert java program from a procedural program to an OO design, and to implement a Java client application	7

S41	Formal Exp.	Academic performance, enjoyment	Results showed positive outcome of the use of PP; with significant improvement in students' academic performance when using PP compared to solo. Performance was measured based on assignment, test, exam scores, and final grade.	Yes	190	Software Design and Construction	Students had to deliver three assignments, each for the duration of three weeks.	7
S42	Formal Exp.	Time spent	The pairs' skill level was found not correlated with how long they spent time on coding.	Yes	38	XP	2 programming tasks: Polynomial and Shuffle puzzle. Data collected in 2 summer term.	5.5
S43	Formal Exp.	Reliability and costs	A pair of developers does not produce more reliable code than a single developer whose code was reviewed. Single programmers developed programs with comparable quality but with fewer cost as compared to pair developers.	Yes	20	XP	4 sessions and one week of project work; students required to complete 2 coding tasks (polynomial and shuffle puzzle)	6.5
S44	Formal Exp.	Time spent	There is no difference in terms of time spent to complete the tasks between the pair programmers and single developer assisted with code review phase However, if program correctness is of no concern, programmer pairs tend to develop programs at higher level of correctness.	Yes	38	XP	Two experiments conducted in two summer lectures; each involved four sessions and a whole week of project work where subjects required to solve two different tasks: Polynomial and shuffle puzzle	6
S45	Formal Exp.	Number of defects	Programmer pairs made as many algorithmic mistakes as solo programmers. However, they perform fewer mistakes for simple problems.	Yes	38	XP	(same as S44)	6
S46	Formal Exp.	Dev. cost (time spent on design, coding and acceptance test), level of correctness	There is no difference in terms of development cost between a pair and a solo implementation if similar level of correctness is concerned.	Yes	18	XP	An experiment composed of four sessions where students had to design and implement a scheduling algorithm and the elevator control of an elevator system.	6.5
S47	Formal Exp.	Total development time, number of defects	There is no difference in development time between PP and XP-based solo programming. Solo programming using PSP seems less efficient than solo programming based on XP. In terms of development time and program size, pair programming is more predictable or stable. Results also indicate that the amount of rework for pair programmers is slightly smaller than solo programmers.	Yes	21	Unk.	Four experiments over a winter semester where students must write four C/C++ programs ranging from 150 to 400 LOC	5.5
S48	Formal Exp.	Mid-term and final exam scores	Using PP greatly improves students' performance. The use of PP as a scaffolding delivery methodology was successful in improving the conceptual understanding of the students. Overall, the class made significant gains in	Yes	10	Internet Prog.	Students initially worked solo but get in paired after the midterm. There was no specific task assigned to the	3

			their conceptual understanding through their performance in final exam.				students and the course did not have a formal lab.	
S49	Formal Exp.	Total development cost (total number of man-hours to measure effort)	All the teams in the pair development group spent less effort to develop the project compared to inspection group.	Only pair	104	SE	Two separate experiments involving students and practitioners; students to develop “TU research resources access control system” over 4 months. Team of developers to develop web application within 5 months.	7
S50	Formal Exp.	Pair effectiveness (measured by communication velocity, productivity and satisfaction	Groups of different personality showed negative correlation between velocity and productivity but perform better when the communication transaction increase.	Only pair	84	SE	Students required to design, code and test two tasks on Coffee Machine Design using Java for a duration of two and a half hours	6.5
S52	Case study	Time (in hour) spent to complete the change request	Programmer pairs worked on the tasks for a significantly shorter duration but with higher cost (i.e double the time). However learning process is faster for the pairs than the individual.	Yes	6	SE	Students had to perform incremental changes to an existing large open source program throughout their course project.	5.5
S53	Case study	Time spent on the task	The pair programmers worked for a significantly shorter duration than the solo programmer.	Yes	12	SE	Implement a simple bowling game application (to record a score) as a course project.	4.5
S54	Formal Exp.	Exam scores	Paired consists of higher confidence students scored higher exam scores than the other group of middle and low confidence, but the difference was not statistically significant.	Only pair	64	Intro. Prog.	Students are required to solve simple problem during the lab hour.	5.5
S55	Formal Exp.	Number of defects found	With PP there was on overall reduction in the number of defects found over time. Thus code becomes more stable as time progresses.	XP Vs TSP	8	Practicum	Implement a spacecraft computer system for duration of 12 weeks.	6
S57	Formal Exp.	Productivity (effort spent on use cases), defects, design quality, knowledge transfer and enjoyment	Paired teams had 29% lower project level productivity than the solo teams (i.e. solo teams finished more use cases). Pair programmers made 8% less defects during development but the system was delivered with higher number of defects compared to solo teams. PP teams had slightly better design quality but no conclusion was made whether paired teams is better in designing software. Paired teams however indicate better knowledge transfer.	Yes	20	J2EE	Students must participate in J2EE training and working 100h for the project. The project included developing, testing and delivering a distributed, multi-player casino system using J2EE technologies. The project effort was fixed to 400hours i.e. 100h per person.	6

S59	Survey/ mixed method	Project scores	Paired students performed very well in their projects with an average score of 98%. The same group of students also performed well in their individual work.	Yes	20	Web Prog.	Develop an e-commerce web site in a course project over a Summer term lasted for 11weeks.	4
S60	Formal Exp.	Development time, productivity and quality	Paired group passed more test cases and their results were more consistent compared to the solo group. Pair groups also completed their assignments 40% to 50% faster.	Yes	41	SE	Students had to deliver 4 programming assignments for the duration of 6 weeks.	6
S61	Formal Exp. Observation	Academic performance	68% of paired students passed the course with a grade C or better compared to 45% of solo students. Differences in final exam score are not statistically significant. The study cannot conclude that PP helps students perform better on exams. Paired students performed better on two of the three programming projects.	Yes	69 (solo) 44 (pair)	CS1	Weekly assignment to be completed in a closed lab during the allotted time. Students were also given programming project.	6.5
S62	Formal Exp.	Success rate, performance on exams (final exam scores)and project scores	An equal or higher percentage of paired students complete the class with a grade of C or better compared to solo programmers. There were no statistical significant differences between the final exam scores for the pairing and non-pairing students. In terms of project score, at the North Carolina State University (NCSU), there were no statistically significant differences in overall project scores between pairing and solo sections. However, the results at University of California Santa Cruz (UCSC) indicate that pair students perform better in project than the solo students.	Yes	Over 1200	CS1	A 3-academic term study in two Universities. At NCSU students to deliver 3 programming projects completed outside lab, and weekly assignments in a closed lab to be completed during the allotted time. At UCSC, 4 or 5 programming projects to be completed outside lab. No specific in lab assignments and no direct supervision of pairing process at UCSC.	6
S64	Formal Exp.	Number of defects	Paired teams showed higher effectiveness in detecting defects for source code and text documents. PP performed better than best-practice inspection concerning defect detection in code documents, whereas paper-based defect detection (UBR) approach performed better for design document.	Yes	41	SE and quality assurance workshop	To perform inspection on a taxi management system within the duration of up to 10 hour	7
S65	Formal Exp.	Effort (Time spent)	Effort for XP is significantly greater than the Side-by-Side programming (SbS). Results indicated that SbS is an interesting alternative to XP-like PP due to less effort overhead. Completion time for SbS was at the level of 60% compared with solo programming, the effort overhead for SbS is as small as 20%. Knowledge about coding spread slower for SbS than for XP. 55% of the	Yes	25	Unk.	To develop Java-based web applications in JSP and Java servlet technology for a duration of 3 months.	5

			subjects preferred PP (both XP and SbS), however 70% of it favour SbS and only 30% favour XP-like PP.					
S71	Formal Exp.	Number and types of problems	Paired students exhibited similar problem distribution as solo students. The number of problems asked by paired students was only 41% of the solo students; thus indicates that paired students require minimal help in solving the tasks. Pairing students are able to resolve more problems on their own, which seems likely to improve their confidence and self-esteem.	Yes (pls chg in the form	15 pairs	Intro. Java Prog.	Students were required to modify existing classes, create classes, and develop some interacting classes in 31 lab exercises. Data of two academic semesters were collected.	4
S72	Formal Exp.	Satisfaction	Overall results suggested that the majority of students are satisfied with PP regardless of pair formation method	Only pair	52	Computer Prog.	Weekly in-lab programming assignment and a course project over a semester.	4.5
S73	Formal Exp.	Code productivity, code design, and satisfaction	Paired students with diverse personality performed significantly better than pairs of similar personality in terms of code productivity & code design. Diverse pairs also obtained better scores in code productivity & code design compared with pairs of opposite personality. Personality differences do not give significant impact on communication, satisfaction, confidence, and compatibility level of the students.	Yes	128	Unk.	Students had to solve 4 programming problems within 90 minutes	5.5
S74	Formal Exp.	Time spent, and the number of test cases passed	Results showed that on average PP was slightly efficient than solo but this was not statistically significant. PP efficiency was directly related to an individual efficiency of students who paired.	Yes	28	Unk.	Two programming problems to be solved in 30 minutes for the first task, and 60 minutes for the second task.	5

Table A.4.4 RQ2 – Studies investigating PP's effectiveness (Qualitative)

Study	Research Design	Purpose of study	Method(s)	Outcome(s)	Sample size	Quality score
S4	Case Studies	Higher productivity, increased confidence, retention in IT careers and positive collaborative experiences.	Semi-structured Interviews, student project retrospective	Conjectures from the case study: 1) Face-to-face meetings appear to be a requirement for timely, high-quality product. 2) Collaboration helps female students build confidence via higher quality product 3) Effective collaboration may help students manage time more effectively.	3	6.5
S10	Survey	To observe the advantage and disadvantages of pair rotation	Questionnaire & observation	Pair rotation helps teachers to obtain multiple peer evaluations on each student; thus reduces pair dysfunctional. Pair rotation allows exposure to more classmates, and ameliorate the partner incompatibility issue	17	5

S14	Field study (exploratory)	To explore factors that may affect the success of pair programming	Participant observation, questionnaires, semi-structured interview and field notes.	Efficiency, enjoyment and perception of learning all showed positive result and favour towards PP.	58	5
S16	Survey	The study focused on PP, pair trading and how they facilitated or hindered software development	Survey Questionnaire	Having a partner helped students complete their assignment (ie. “pair-relaying” factor). Students who responded negatively cited difficulty in scheduling and working with a weak partner. Almost all respondent agreed that PP increased reliability and confidence.	22	3.5
S24	Grounded Theory	Academic performance (GPA score, assignment score, and final scores).	Interview and project retrospective questionnaire	Pair groups did not do well as the solo groups (based on students’ grade on the final project). Schedule mismatch and bad pairing experience is the enemy of effective PP.	15	6
S51	Case Studies	To find out the effects of collaborative pedagogy on student perceptions	Interviews, retrospectives	PP helps increased productivity and faster learning, increased confidence and quality of work. Collaborative environments are viable pedagogical alternatives that can improve CS education.	6	6.5
S56	Survey	To understand the perceived benefits of PP when integrated with written report.	Surveys, Written report	PP helps increased students’ confidence, increased understanding on the project solution, and helps them being more efficient in debugging. The written report helped students to understand the processes involved during the PP session. 48% of the students’ responses that the main disadvantage of PP was to find time for a meeting.	293	5
S58	Case studies	To investigate whether PP indeed improves learning in a CS classroom and looking at the best way to implement it.	Observation & interviews	The interview result showed that nearly all students preferred to work together on programming assignments. PP works best when the skill levels gap of the pair is not too broad.	1	3.5
S66	Survey	To investigate the use of PP in an introductory course where issues like non-residential campus and working students were addressed.	Surveys, and reflective paragraph	Students reported positive experience when using PP despite of scheduling challenges that they faced. (Students strongly indicated positive attitudes towards working with a partner) Their results showed that PP can indeed be successful at a commuter campus.	80	4.5
S69	Case studies	To investigate how do students define, experience and value PP.	Semi-structured interviews	The findings from this study appear to contradict with existing literature, that students understand their work better when program solo. Students in this study felt that PP prepared them for SP, and PP is a good way to learn programming.	11	5
S70	Case studies	To determine if and in what ways PP helps in attract and retain students in IT disciplines. This study focuses on female and minorities students.	Semi-structured interviews	Overall findings suggested that PP effectively create the collaborative environment desirable for students.	6	5

Table A.4.5 RQ3 – Quality metrics used in quantitative studies

Study	Type of Study	Quality Measure(s)	Summary of findings	Course(s) involved	Pair Vs. Solo?	Sample size
S1	Formal Exp.	Using quality metric defined in ISO/IEC 9126 to measure functionality, usability, portability and maintainability of Data Flow Diagrams, Relational Data Bases, and Functional Interface Diagrams.	Pair programmers produce significantly better design in terms of functionality, usability and portability. However, for more complex requirements, there was no significant difference in any quality characteristics between paired and solo designers.	Unk.	Yes	150
S2	Formal Exp.	Programming score (in scale of 1 – 125, where 1 is the lowest score, and 125 is the highest score) rated based on grading scheme developed by a faculty expert.	Paired students produce higher quality software. Task complexity does not affect the quality of software produced by either pair or solo programmers.	Information System course	Yes	120 (3 acad. Sem.)
S5	Formal Exp.	Based on the Chidamber and Kemerer (CK) metric suit. Measuring quality was also done using expert judgement to inspect programs and their quality	Paired programmers produce code with higher quality compared to soloist. In terms of code complexity, there was a trend that paired teams produce less complex programs. According to expert judgement, code written by paired teams was rated a little bit better. Their readability and understandability were somewhat higher.	Software-praktikum course (SoPra)	Yes	165 (2 acad. Sem.)
S9	Formal Exp.	Assignment score: 0 (incorrect), 0.5 (neither incorrect nor completely correct), and 1 (correct)	The quality achieved with pair designing is significantly greater than the solo designing.	Unk.	Yes	70
S13	Survey & Formal Exp.	Code quality was measured based on output correctness, documentation, programming style, correct use of objects and interface design.	Quality of code does not determined by personality traits of the pair developers. (The measuring was done by a single person to ensure consistency).	CS1	Only pair	118
S20	Formal Exp.	Project scores	There is no significant difference in performance between pair and solo students in their course project.	Adv. Microprocessor	Yes	101
S21	Formal Exp.	Number of features correctly implemented, LOC and the cyclomatic number (CCN) of the program (Objective measure). Use of meaningful identifiers, well-organized methods, appropriate indentation and whitespace, and use of Booleans (Subjective measures).	Paired students perform better in terms of the number of features implemented. There was mixed-result in terms of complexity and program size of pair and solo students. There is an evident trend that the length and complexity of programs produced by pairs will increase as the difficulty of the assignments increased.	CS1	Yes	150 approx.
S25	Case study	LOC (without comment), Comment Ratio (CR), and Coupling Factor (CF)	LOC for PP teams is slightly lower than non-PP teams. PP teams had slightly lower CR and CF. So, lower quality for PP teams in terms of CR, but a higher quality if based on CF.	Software-Praktikum	Yes	24 (4 teams)
S30	Formal Exp	Design scores	Paired students produced better quality design than solo students.	Prog. languages, Internet prog., Database prog. and Systems prog. courses	Yes	7

S33	Formal Exp	Quality was measured based on number of passed tests.	Pair programmer can produce software of better quality when the pair is new to a programming problem.	Agile Software Dev. and XP	Yes	40
S34	Formal Exp	The external code quality was measured by a number of acceptance tests passed (NATP)	Code quality was significantly affected by the software development approach (i.e. Classical/ test driven development on solo/pair programming) Classical approach (pair & solo) achieved better results compared to a test-driven approach. Test-driven development decreases the external code quality when used by either pair or solo students. There was no difference in the external code quality when PP was used instead of solo programming.	Programming in Java (PIJ)	Yes	188
S35	Formal Exp	Quality of the object-oriented design was measured using the Martin's package level dependency metrics. The data was collected using *aopmetrics* tool.	Results indicate imperfect package design regardless of software development method used. The study found that package level design quality indicators were not significantly affected by development method (pair or solo).	Programming in Java (PIJ)	Yes	122
S36	Formal Exp	The quality focus was based on thoroughness and fault-finding effectiveness of unit test suites	Software development approach (pair or solo) does not give significant impact on both thoroughness and fault-finding effectiveness of unit test.	Programming in Java (PIJ)	Yes	98
S38	Formal Exp	Functionality and style (measured objectively). "holistic" (measured subjectively)	Programs written by pairs obtained higher average rankings in the holistic evaluation. Individual programs were better in terms of functionality and style rankings, but provided poor design. Overall evaluation showed that the programs produced by paired students are significantly better than the programs produced by solo students for the same or comparable assignments.	Advanced Programming, and Abstract Data Types	Yes	100 approx.
S39	Formal Exp	Average programming scores	Average programming score for paired students was significantly better than solo students (86.6% Vs 68.1%).	CS1	Yes	555
S44	Formal Exp.	Number of test cases passed	Pair programmers and solo developer assisted with code review phase are interchangeable in terms of development cost (ie. Time spent). However, if a similar level of program correctness is of no concern, programmer pairs tend to develop programs at higher level of correctness.	XP	Yes	38
S45	Formal Exp	Number of defects/errors (classified into specification, expression and algorithm). Defect classification was done by experts.	Both pair and solo programmers make similar algorithmic errors. However, pair programmers made fewer mistakes for simple problems, compared to solo programmers.	XP	Yes	38 (2 acad. Sem.)
S46	Formal Exp.	Acceptance-test was used to identify the level of correctness of developed programs	Pair programmers tend to develop programs with fewer numbers of failures. However there is no significant difference in terms of number of failures after coding between pair and solo programmers due to a small sample size used.	XP	Yes	18

S47	Formal Exp	Total LOC per hour per person and number of errors uncovered during acceptance testing	A version of XP for a single programmer was found the most efficient in terms of LOC per hour per person (more code written in less time). It appears that solo and PP achieve more or less the same performance. PP leads to more stable solutions compared to others. PP also performs slightly better in terms of number of errors uncovered.	Unk.	Yes	21
S49	Formal Exp	Product quality was based on number of un-passed test cases and the number of incomplete requirements.	There is no statistical significant difference regarding the un-passed test cases and incomplete requirements between the paired teams and software inspection teams.	SE	Only pair	104
S52	Case studies & surveys	Based on the completion of change requests.	Pair programmers completed the change requests faster than solo programmers. They also did most of the job correctly with slightly fewer lines of code; they use meaningful variables thus produce higher quality code compared to solo programmer.	SE	Yes	6
S53	Case studies	Lines of Code (LOC), number of methods, number of passed test cases.	Average LOC for programs developed by pair programmers is higher compared to solo programmers. Solo programmers write code in a single class whereas pair programmers used multiple classes. Pairs also created much more class members (methods) and this indicates higher cohesion of code. Program written by pairs passed more test cases than programs produced by individuals.	SE	Yes	12
S55	Formal Exp	Number of defects	Programs developed by paired programmers contained less number of defects. Thus, codes written by pairs were more stable.	Practicum	Yes	8
S57	Formal Exp	Design quality was measured on the method level. Non Comment Lines of Code (NCLOC) per method was used to characterize the method size, McCabe's cyclomatic complexity was used to describe method's complexity. Number of parameters was used to tell how much information is passed to a method.	Paired teams had slightly better design quality based on the method size and complexity metrics. The differences between pair and solo teams depend on the metric used and the metrics may be affected by the size of the analyzed code. Thus, there was no evidence that paired teams produce better software design than solo teams.	J2EE	Yes	20
S59	Surveys	Project scores	The effects of "pair pressure" through PP practice bring positive effect towards product quality. Product was delivered in timely manner. (The average grade on all 80 assignments was 98%).	Web Programming	Only pair	20
S60	Formal Exp.	Percentage of test cases passed	Programs developed by pair programmers passed more automated test cases than the programs developed by solo programmers. The percentage of test cases passed for all programs were more consistent for pair programmers.	SE	Yes	41
S62	Formal Exp.	Project score	The result was contradicted between studies at University of California Santa Cruz (UCSC) and North Carolina State University (NCSU). At NCSU, there is no difference in project	CS1	Yes	> 1200 students in 2

			scores between solo and pairing sections. However, at UCSC, paired students scored better marks in their projects compared to solo students.			different Univ.
S64	Formal Exp.	Number of defects	Paired teams showed higher effectiveness in detecting defects for source code and text documents. PP performed better than best-practice inspection concerning defect detection in code documents; paper-based defect detection (UBR) approach performed better for design document.	SE & Quality assurance workshop	Yes	41
S68	Formal Exp.	External code quality (measured by number of acceptance tests passed or NATP)	External code quality was correlated with the feelgood factor and programming experience. Both factors (feelgood and years of experience) may be the drivers for the external code quality	Programming in Java (PIJ)	Only pair	132
S73	Formal Exp.	Code design (measured subjectively by professional judges)	Code design quality was higher for PP teams than in solo programming teams.	Unk.	Yes	128
S74	Formal Exp.	Number of test cases passed	Pair developers produced programs of higher quality than program produced by solo developers in terms of number of test cases passed	Unk.	Yes	28

Table A.4.6 RQ3 – Quality metrics used in qualitative studies

Study	Type of Study	Quality Measure(s)	Summary of findings	Course(s) /students involved	Pair Vs. Solo?	Sample size
S16	Survey	Code quality was defined as code that contains fewer errors and fulfil the requirements specifications (students rated the code themselves, against code they wrote when worked solo)	PP helps increase students' confidence in completing the program. Students also perceived that code quality was greater when pairing compared with working solo.	CS2	Only pair	22

APPENDIX B: Experiment Resources

Appendix B.1 Ethics Approval Letter

Office of the Vice-Chancellor
Ethics and Biological Safety Secretariat

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Level 3, 76 Symonds Street
Telephone: 64 9 373 7599
Extension: 83711 / 87830
Facsimile: 64 9 3737432

The University of Auckland
Private Bag 92019
Auckland Mail Centre
Auckland 1142
New Zealand

UNIVERSITY OF AUCKLAND
HUMAN PARTICIPANTS ETHICS COMMITTEE

11 September, 2008

MEMORANDUM TO:

Norsaremah Salleh

Computer Science

Re: Change to application

I wish to advise you that the Committee met on 10 September , 2008 and reviewed the request for change to your application titled "Improving the effectiveness of pair pairing as a Pedagogical Tool for Computer Science/Software Engineering Education: an Investigation of Pair Compatibility" (Our Ref. 2008 / 291).

The Committee approved the change.

If the project changes significantly you are required to resubmit your appplication to the Committee for further consideration.

In order that an up-to-date record can be maintained, it would be appreciated if you could notify the Committee once your project is completed.

Please contact the Chairperson if you have any specific queries relating to your application. The Chair and the members of the Committee would be most happy to discuss general matters relating to ethics provisions if you wish to do so.

Lana Lon
Executive Secretary
University of Auckland Human Participants Ethics Committee

c.c. Head of Department / School, Computer Science

Norsaremah Salleh

13G, 32 Eden Cresent
Auckladn CBD

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Building 303
Level 3, 38 Princes Street
Auckland 1142, New Zealand
Telephone 64 9 373 7599 ext.85857
Facsimile 64 9 373 7453
Email: office@cs.auckland.ac.nz
www.cs.auckland.ac.nz

The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

Appendix B.2 Participant Information Sheet

Improving the Effectiveness of Pair Programming as a Pedagogical tool for Computer Science/Software Engineering Education: An Investigation of Pair Compatibility

Participant Information Sheet (Students)

My name is Norsaremah Salleh and I am a doctoral student at the Department of Computer Science, The University of Auckland. I am conducting a pair programming study which will take place during some of the tutorial sessions of the COMPSCI 101 course. You are invited to participate in this research and I would appreciate any assistance you can offer.

Pair Programming (PP) is a technique where two people sitting side-by-side, working together on the same algorithm, design, code or test. One of them is the "driver", who is responsible for designing, typing the code, and who has control over the resources such as computer, mouse and keyboard; the other is known as "navigator" or "observer", and has the responsibility of observing how the driver works, to detect errors made by the driver and offer ideas in solving a problem. Throughout their work, pairs alternate their roles every 20 minutes. The research I am conducting aims to investigate the ways to improve the effectiveness of PP as a pedagogical tool focusing on the impact of psychosocial factors towards the practice. In particular, we seek to understand how the personality and gender combination affect the results of using PP. As part of the research, participants will be requested to fill out an online personality test, which will consume 10-20 minutes of your time. The test will identify your personality based on the Big-Five factor, and please note that your personality profile will be kept confidential to me as the researcher. During the tutorial session, you will be required to work in pairs. Participants will be paired randomly using the personality results as basis for pair formation. At the end of the tutorial session, participants are required to fill out a short-item questionnaire. The questionnaire and the personality data will be securely stored at the Department of Computer Science, University of Auckland, for six (6) years, after which they will be destroyed. For publishing and writing purposes, care will be taken to preserve the confidentiality of the participants. As also part of the research, I will ask participants to make their assignments, and exam grades available for aggregated analysis in this research.

Participation in this research is on a voluntary basis. You can be assured that neither your grades nor academic relationships with the department staff members will be affected by either refusal or agreement to participate. You have the right to withdraw yourself from the study at any time. If you consent to participate, you also have the right to withdraw the information you have already provided by the 24th October 2009. Since the study will take place in tutorial sessions, non-participants will be informed that they will be working individually. Please also note that the anonymity of non-participants will be preserved by the

researcher. Participants are expected to write their UPI on the questionnaire form as the information is pertinent for this research. If throughout the session, participants feel discomfort working with his/her partner, the student will be advised to work individually for the rest of the session.

This research is funded by the Ministry of Higher Education, Malaysia. If you have any queries regarding this study, please do not hesitate to contact me. You can email me at: norsaremah@gmail.com OR nsal017@aucklanduni.ac.nz. Alternatively, you may phone me at 09 373 7599 ext 87625. You may also contact my main supervisor Associate Professor Emilia Mendes at emilia@cs.auckland.ac.nz or 09 373 7599 ext. 86137 or the Head of Department, Professor Gill Dobbie at gill@cs.auckland.ac.nz or 09 373 7599 ext. 83949.

For any queries regarding ethical concerns you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Office of the Vice Chancellor, Private Bag 92019, Auckland 1142. Telephone 09 373-7599 extn. 83711.

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 10/09/2008 for 3 years on 10/09/2008 to 10/09/2011

Reference Number 2008/291

Instructions for the Personality Test

1. Open your Internet browser (Firefox/ Internet Explorer)
2. Type this URL: http://www.personal.psu.edu/j5j/IPIP/
3. Please click on the short version of IPIP-NEO
4. Please enter your UPI as the nickname and then complete the test.
5. Please save the results in an HTML file or Document file using your UPI as the file name. (eg: nsal017.htm)
6. Email the results to compsc101@gmail.com

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

Building 303
Level 3, 38 Princes Street
Auckland 1142, New Zealand
Telephone 64 9 373 7599 ext.85857
Facsimile 64 9 373 7453
Email: office@cs.auckland.ac.nz
www.cs.auckland.ac.nz

The University of Auckland
Private Bag 92019
Auckland 1142
New Zealand

Appendix B.3 Consent Form

Improving the Effectiveness of Pair Programming as a Pedagogical tool for Computer Science/Software Engineering Education: An Investigation of Pair

This Consent Form will be stored for six (6) years.

Consent Form (Students)

I have read and understood the Participant Information Sheet. I understand the nature of the research and why I have been selected to participate in this research. I have been given the opportunity to ask questions about the study.

- I understand that the information about my personality profile will be gathered.
- I understand that I will need to fill up a short-item questionnaire at the end of the tutorial session if I choose to participate in this study.
- I understand that only the researcher and her main supervisor will have access to questionnaire form and the personality data.
- I have been informed that the information that I provide will be kept safely at the University of Auckland, held for analysis for 6 years, after which they will be destroyed.
- I understand that my grades will be used for aggregated analysis in the research.
- **I understand that neither my grades nor academic relationship with any department staff members will be affected by either refusal or agreement to participate.**
- I understand that I will not be directly identified as an individual source of the information that I provide.
- I understand that I have the right to withdraw from this study at any time.
- I understand that I have the right to withdraw the information that I provide at any time before 24th October, 2009.

I agree to take part in this study under the terms and conditions provided to me.

Name:

Signature & Date:

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 10/09/2008 for 3 years on 10/09/2008 to 10/09/2011
Reference Number 2008/291

Appendix B.4 Pair Programming Questionnaire

Please enter your UPI (e.g. nsal017): _____Computer ID (e.g.A1) _____

Instruction:
Please answer the following questions without discussing with your partner. **All responses will be treated in the strictest confidence**. For Q1 until Q7, please tick your answer using the following scale: *1:Strongly Disagree (SD) 2:Disagree (D) 3:Neither Agree Nor Disagree (N) 4: Agree (A) 5:Strongly Agree (SA)*

		1 (SD)	2 (D)	3 (N)	4 (A)	5 (SA)
Q1	I felt that working with this partner was a productive experience.					
Q2	I enjoyed working with my partner.					
Q3	My motivation level increased when working with my partner.					
Q4	I understood the topic better when working with my partner.					
Q5	My level of confidence in solving the exercises increased when working with my partner.					
Q6	I felt it was a waste of time working with my partner.					

Q7. Please rate how satisfied are you working with your partner. (Circle only one).

Not satisfied at all

1 2 3 4

Q8. How do you rate your level of confidence solving the exercises with your partner? (Circle only one)

Very Low

1 2 3 4

Q9. Comments:

_____

_____

Thank you for your time!

Please let us know if you have any queries about this questionnaire or the study we are conducting. Questions or concerns can either be directed to the researcher, Norsaremah Salleh (nsal017@aucklanduni.ac.nz) or to Assoc. Prof. Emilia Mendes (emilia@cs.auckland.ac.nz), Dept. of Computer Science.

Appendix B.5 Personality Test and Demographic Survey

Please enter your UPI (Net ID): _____

All responses will be treated in the strictest confidence.

Section A: Demographic Information

A1: Date of Birth (dd/mm/yyyy):____/___/_____

A2: Gender (please tick):

☐ Male ☐ Female

A3: Ethnicity/Race (please tick):

☐ NZ/ Pakeha	☐ Chinese
☐ Maori	☐ Indian
☐ Pacific Islander	☐ Others (please specify):_____

A4: Please state the number of years of your working experience related to IT industry: _____ years

A5: On a scale from 1 – 5, how do you rate your programming competency level? (Please tick)

☐ 1. Very Poor ☐ 2.Poor ☐ 3.Fair ☐ 4.Good ☐ 5.Outstanding

A6: Have you ever experience pair programming or any collaborative work before? (Please underline) YES / NO

A7: Is English your first language? (Please underline) YES / NO

Section B: Personality Test

Instructions:

7. Open your Internet browser (Firefox/ Internet Explorer)
8. Type this URL: http://www.personal.psu.edu/j5j/IPIP/
9. Please click on the **short version** of IPIP-NEO
10. Please enter your UPI as the nickname and then complete the test.
11. Please save the results in an HTML file or word document file using your UPI as the file name. (eg: nsal017.htm)
12. Email the results to **compsc101@gmail.com**

Thank you for your time!

APPENDIX C: PALLOC Resources

Appenxix C.1 PALLOC Database Structure

Table 1: Students

Attribute	Type
Stud_UPI	varchar(8)
First name	char(40)
Surname	char(30)
Gender	char(1)
Extraversion	tinyint (4)
Agreeableness	tinyint (4)
Conscientiousness	tinyint (4)
Neuroticism	tinyint (4)
Openness	tinyint (4)
Personality_group	tinyint (4)

Primary Key: Stud_UPI
Personality group category: 0= low conscientiousness; 1=medium conscientiousness; 2=high conscientiousness; 3=unknown

Table 2: Paired

Attribute	Type
Stud_UPI	varchar(8)
Partners_UPI	varchar(8)
Lab_ID	Smallint(6)
Pair_type*	tinyint (4)

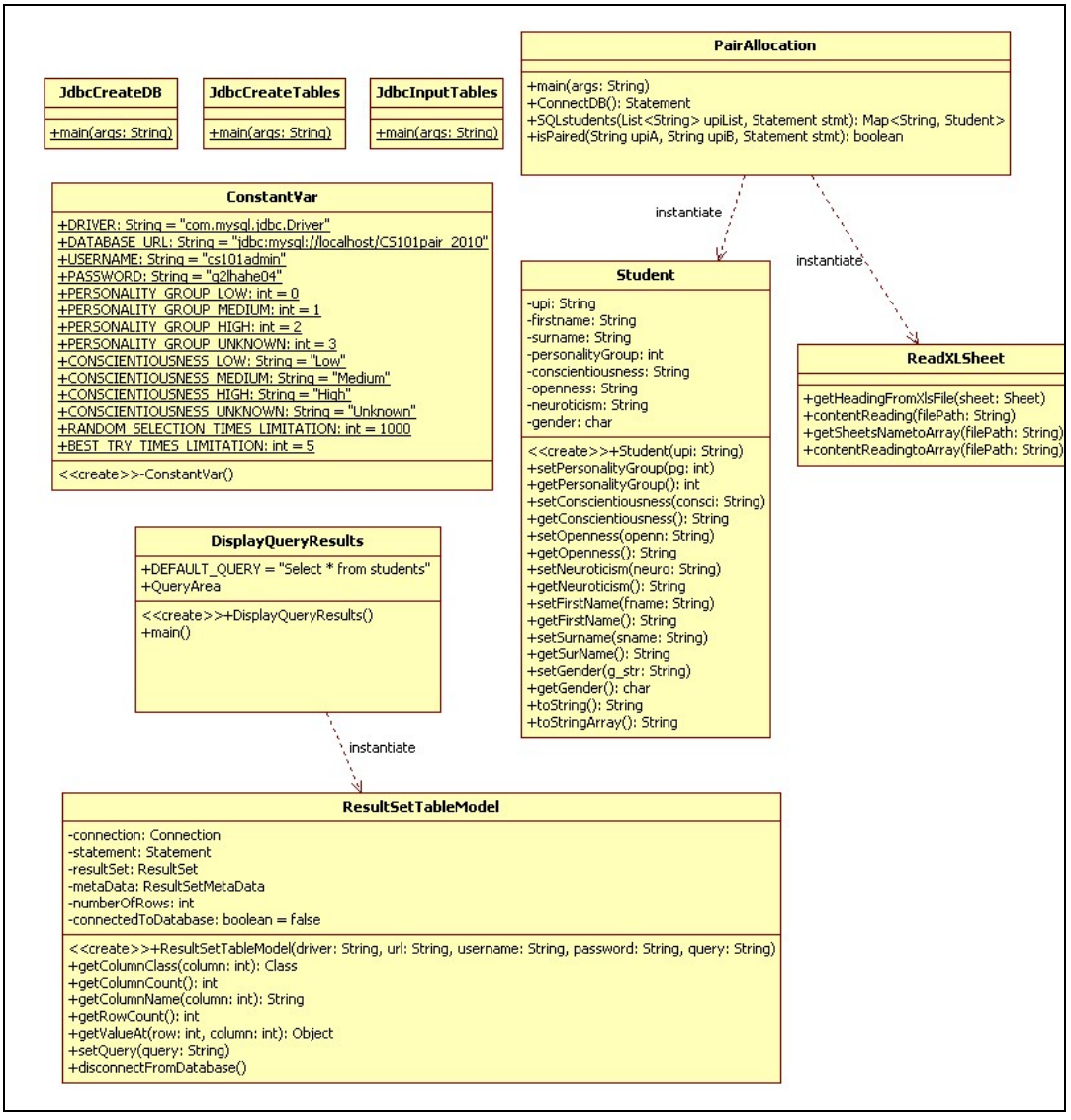
Primary Key: Stud_UPI and Partners_UPI
*Pair_type – The value depends on the level of conscientiousness of both students (e.g. if both students are having low conscientiousness means they are in “low conscientiousness” group, hence set pair_type to zero (0))

Table 3: Lab Session

Attribute	Type
Lab_ID	smallint(6)
Lab_Date	Date
Lab_Session	Varchar(15)

Primary Key: Lab_ID

Appendix C.2 Overview of Design Model (PALLOC)



APPENDIX D: PERSONALITY INSTRUMENT

Instructions for Completing the IPIP-NEO Short Form

The following pages contain phrases describing people's behaviors. Please use the rating scale next to each phrase to describe how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence.

Your UPI:		Gender	☐ Male ☐ Female
Age:		Country of origin:	

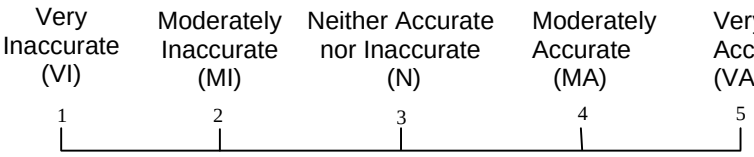
Please tick your answer using the following scale:

Very Inaccurate (VI)	Moderately Inaccurate (MI)	Neither Accurate nor Inaccurate (N)	Moderately Accurate (MA)	Very Accurate (VA)
1	2	3	4	5

		1 (VI)	2 (MI)	3 (N)	4 (MA)	5 (VA)
1	Worry about things.					
	Make friends easily.					
3	Have a vivid imagination.					
4	Trust others.					
5	Complete tasks successfully.					
6	Get angry easily.					
7	Love large parties.					
8	Believe in the importance of art.					
9	Use others for my own ends.					
10	Like to tidy up.					
11	Often feel blue.					
12	Take charge.					
13	Experience my emotions intensely					
14	Love to help others					
15	Keep my promises.					
16	Find it difficult to approach others.					
17	Am always busy.					
18	Prefer variety to routine.					
19	Love a good fight.					
20	Work hard.					
21	Go on binges.					
22	Love excitement.					
23	Love to read challenging material.					
24	Believe that I am better than others.					
25	Am always prepared.					
26	Panic easily.					
27	Radiate joy.					
28	Tend to vote for liberal political candidates.					

Very Inaccurate (VI)	Moderately Inaccurate (MI)	Neither Accurate nor Inaccurate (N)	Moderately Accurate (MA)	Very Accurate (VA)

		1 (VI)	2 (MI)	3 (N)	4 (MA)	5 (VA)
29	Sympathize with the homeless.					
30	Jump into things without thinking.					
31	Fear for the worst.					
32	Feel comfortable around people.					
33	Enjoy wild flights of fantasy.					
34	Believe that others have good intentions.					
35	Excel in what I do.					
36	Get irritated easily.					
37	Talk to a lot of different people at parties.					
38	See beauty in things that others might not notice.					
39	Cheat to get ahead.					
40	Often forget to put things back in their proper place.					
41	Dislike myself.					
42	Try to lead others.					
43	Feel others' emotions.					
44	Am concerned about others.					
45	Tell the truth.					
46	Am afraid to draw attention to myself.					
47	Am always on the go.					
48	Prefer to stick with things that I know.					
49	Yell at people.					
50	Do more than what's expected of me.					
51	Rarely overindulge.					
52	Seek adventure.					
53	Avoid philosophical discussions.					
54	Think highly of myself.					
55	Carry out my plans.					
56	Become overwhelmed by events.					
57	Have a lot of fun.					
58	Believe that there is no absolute right or wrong.					
59	Feel sympathy for those who are worse off than myself.					
60	Make rash decisions.					
61	Am afraid of many things					
62	Avoid contacts with others.					
63	Love to daydream.					
64	Trust what people say.					
65	Handle tasks smoothly.					
66	Lose my temper.					
67	Prefer to be alone.					
68	Do not like poetry.					
69	Take advantage of others.					
70	Leave a mess in my room.					
71	Am often down in the dumps.					
72	Take control of things.					
73	Rarely notice my emotional reactions.					



		1 (VI)	2 (MI)	3 (N)	4 (MA)	5 (VA)
74	Am indifferent to the feelings of others.					
75	Break rules.					
76	Only feel comfortable with friends.					
77	Do a lot in my spare time.					
78	Dislike changes.					
79	Insult people.					
80	Do just enough work to get by.					
81	Easily resist temptations.					
82	Enjoy being reckless.					
83	Have difficulty understanding abstract ideas.					
84	Have a high opinion of myself.					
85	Waste my time.					
86	Feel that I'm unable to deal with things.					
87	Love life.					
88	Tend to vote for conservative political candidates.					
89	Am not interested in other people's problems.					
90	Rush into things.					
91	Get stressed out easily.					
92	Keep others at a distance.					
93	Like to get lost in thought.					
94	Distrust people.					
95	Know how to get things done.					
96	Am not easily annoyed.					
97	Avoid crowds.					
98	Do not enjoy going to art museums.					
99	Obstruct others' plans.					
100	Leave my belongings around.					
101	Feel comfortable with myself.					
102	Wait for others to lead the way.					
103	Don't understand people who get emotional.					
104	Take no time for others.					
105	Break my promises.					
106	Am not bothered by difficult social situations.					
107	Like to take it easy.					
108	Am attached to conventional ways.					
109	Get back at others.					
110	Put little time and effort into my work.					
111	Am able to control my cravings.					
112	Act wild and crazy.					
113	Am not interested in theoretical discussions.					
114	Boast about my virtues.					
115	Have difficulty starting tasks.					
116	Remain calm under pressure.					
117	Look at the bright side of life.					
118	Believe that we should be tough on crime.					
119	Try not to think about the needy.					
120	Act without thinking.					

REFERENCES

Abraham, W.T., & Russell, D.W. (2008). Statistical power analysis in psychological research. *Social and Personality Psychology Compass, 2*(1), 283-301.

Ackerman, P.L., & Heggestad, E.D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219-245.

Acuna, S.T., Gomez, M., & Juristo, N. (2009). How do personality, team process and task characteristics relate to job satisfaction and software quality? *Information and Software Technology, 51*, 627-639.

Ally, M., Darroch, F., & Toleman, M. (2005). A framework for understanding the factors influencing pair programming success. In *Extreme Programming and Agile Processes in Software Engineering, Proceedings* (pp. 82-91). Berlin: Springer-Verlag Berlin.

Arisholm, B., Gallis, H., Dyba, T., & Sjoberg, D.I.K. (2007). Evaluating pair programming with respect to system complexity and programmer expertise. *IEEE Transactions on Software Engineering, 33*(2), 65-86.

Baheti, P., Gehringer, E., & Stotts, D. (2002). Exploring the efficacy of distributed pair programming. *Extreme Programming and Agile Methods - XP/Agile Universe 2002, LNCS 2418, Springer-Verlag*, 208-220.

Baron, R.M., & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Barrick, M.R., & Mount, M.K. (1991). The big five personality dimensions and job performance: A Meta-Analysis. *Personality Psychology, 44*, 1-26.

Barrick, M.R., Mount, M.K., & Judge, T.A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *Personality and Performance, 9*(1-2), 9-30.

Barrick, M.R., Stewart, G.L., Neubert, M.J., & Mount, M.K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology, 83*(3), 377 - 391.

Barry, B., & Stewart, G.L. (1997). Composition, process, and performance in self-managed groups: The role of personality. *Journal of Applied Psychology, 82*(1), 62-78.

Basili, V.R., & Rombach, H.D. (1988). The TAME Project: Towards Improvement-Oriented Software Environments. *IEEE Transactions on Software Engineering, 14*(6).

Basili, V.R., Shull, F., & Lanubile, F. (1999). Building knowledge through families of experiments. *IEEE Transaction on Software Engineering, 25*(4), 456-473.

Bax, L., Yu, L.-M., Ikeda, N., Tsuruta, H., & Moons, K.G. (2006). Development and validation of MIX: Comprehensive free software for meta-analysis of causal research data. *BMC Medical Research Methodology, 6*(50).

Bax, L., Yu, L.-M., Ikeda, N., Tsuruta, H., & Moons, K.G. (2008). Comprehensive free software for meta-analysis of causal research data version 1.7. Retrieved from http://mix-for-meta-analysis.info

Beck, K. (1999). *Extreme Programming Explained: Embrace Change* (2nd ed.). Boston, US: Addison-Wesley.

Begel, A., & Nagappan, N. (2008). Pair programming: What's in it for me? *Proceedings of the 2nd Int'l Symp. Empirical Software Engineering & Measurement*, 120-128.

Bell, S.T. (2007). Deep-level composition variables as predictors of team performance: A Meta-Analysis. *Journal of Applied Psychology, 92*(3), 595-615.

Berenson, S.B., Slaten, K.M., Williams, L., & Ho, C.-w. (2004). Voices of women in a software engineering course: Reflections on collaboration. *Journal of Educational Resources in Computing (JERIC), 4*(1).

Bjork, R., & Druckman, D. (1991). *In the Mind's Eye: Enhancing Human Performance*. Washington, DC: National Academy Press.

Blickle, G. (1996). Personality traits, learning strategies, and performance. *European Journal of Personality, 10*, 337-352.

Boekaerts, M. (1996). Personality and the psyhocology of learning. *European Journal of Personality, 10*, 377-404.

Bowers, C.A., Pharmer, J.A., & Salas, E. (2000). When member homogeneity is needed in work teams: A Meta-Analysis. *Small Group Research, 31*(3), 305 - 327.

Boyle, G.J. (1995). Myers-Briggs type indicator (MBTI): some psychometric limitations. *Australian Psychologist, 30*(1), 71-74.

Bradley, J.H., & Hebert, F.J. (1997). The effect of personality type on team performance. *Journal of Management Development, 16*(5), 337-353.

Braught, G., Eby, L.M., & Wahls, T. (2008). The effects of pair programming on individual programming skill. *Proceedings of the ACM Technical Symposium on Computer Science Education (SIGCSE'08)*, 200-204.

Braught, G., MacCormick, J., & Wahls, T. (2010). *The benefits of pairing by ability.* Paper presented at the Proceedings of the 41st ACM Technical Symposium on Computer Science Education (SIGCSE'10).

Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems & Software, 80*, 571-583.

Brereton, P., Turner, M., & Kaur, R. (2009). Pair programming as a teaching tool: A student review of empirical studies. *Proceedings of the 22nd Conference on Software Engineering Education and Training*, 240-247.

Britt, C.L., & Weisburd, D. (2010). Statistical power. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of Quantitative Criminology*: Springer Science+Business Media.

Buchanan, T., Johnson, J.A., & Goldberg, L.R. (2005). Implementing a five-factor personality inventory for use on the Internet. *Journal of Psychological Assessment 2005, 21*(2), 115-127.

Budgen, D., Kitchenham, B., Charters, S., Turner, M., Brereton, P., & Linkman, S. (2007). Preliminary results of a study of the completeness and clarity of structured abstracts. *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering (EASE 2007)*, 64 – 72.

Burch, G.S.J., & Anderson, N. (2008). Personality as a predictor of work-related behavior and performance: Recent advances and directions for future research. In G. P. Hodgkinson & J. K. Ford (Eds.), *International Review of Industrial and Organizational Psychology* (pp. 261-305): John Wiley & Sons, Ltd.

Burger, J.M. (1993). *Personality* (3rd. ed.). California: Brooks/Cole Publishing Co.

Busato, V.V., Prins, F.J., Elshout, J.J., & Hamaker, C. (2000). Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. *Personality and Individual Differences, 29*(6), 1057-1068.

Canfora, G., Cimitile, A., Garcia, F., Piattini, M., & Visaggio, C.A. (2007). Evaluating performances of pair designing in industry. *Journal of Systems and Software, 80*(8), 1317-1327.

Canfora, G., Cimitile, A., & Visaggio, C.A. (2005). Empirical study on the productivity of the pair programming. *Proc. 6th International Conference on Extreme Programming and Agile Processes in Software Engineering (XP 2005), LNCS 3556*, 92-99.

Capretz, L.F. (2002). Implications of MBTI in software engineering education. *SIGCSE Bulletin, 34*(4), 134-137.

Carver, R.H., & Nash, J.G. (2006). *Doing data analysis with SPSS version 14*. Belmont, CA: Thomson Brooks/Cole.

Cegielski, C.G., & Hall, D.J. (2006). What makes a good programmer? *Communications of the ACM, 49*(10), 73-75.

Chamorro-Premuzic, T., & Furnham, A. (2003a). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality, 37*, 319-338.

Chamorro-Premuzic, T., & Furnham, A. (2003b). Personality traits and academic examination performance. *European Journal of Personality, 17*, 237-250.

Chamorro-Premuzic, T., & Furnham, A. (2008). Personality, intelligence and approaches to learning as predictors of academic performance. *Personality and Individual Differences, 44*(7), 1596 - 1603.

Chao, J., & Atli, G. (2006). Critical personality traits in successful pair programming. *AGILE'06, IEEE Computer Society*, 89-93.

Chaparro, E.A. (2005). Factors affecting the perceived effectiveness of pair programming in higher education. *17th Workshop of the Psychology of Programming Interest Group, Sussex University*, 5-18.

Chigona, W., & Pollock, M. (2008). Pair programming for Information Systems students new to programming: Students' experiences and teachers' challenges. *Portland International Conference on Management of Engineering & Technology (PICMET*

2008), 1587-1594.
Choi, K.S. (2004). *A discovery and analysis of influencing factors of pair programming.* Unpublished Ph.D. Dissertation, New Jersey Institute of Technology, USA.
Choi, K.S., Deek, F.P., & Im, I. (2008). Exploring the underlying aspects of pair programming: The impact of personality. *Information and Software Technology, 50*(11), 1114-1126
Cicirello, V.A. (2009). On self-selected pairing in CS1: Who pairs with whom? *Journal of Computing Sciences in Colleges, 24*(6), 43-49.
Cliburn, D.C. (2003). Experiences with pair programming at a small college. *Journal of Computing Sciences in Colleges, 19*(1), 20-29.
Cockburn, A. (2002). *Agile Software Development*. Boston, MA.: Addison-Wesley Longman Publishing Co. Inc.
Cockburn, A., & Williams, L. (2001). The Costs and Benefits of Pair Programming. In *Extreme Programming Examined* (pp. 223 - 243). Boston, MA: Addison-Wesley Longman Publishing Co., Inc.
Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Hillsdale, N.J.: L. Erlbaum Associates.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.
Comrey, A.L., & Staats, C.K. (1955). Group performance in a cognitive task. *Journal of Applied Psychology, 39*, 354-356.
Conard, M.A. (2006). Aptitude is not enough: How personality and behavior predict academic performance. *Journal of Research in Personality, 40*, 339 - 346.
Conn, S.R., & Rieke, M.L. (1994). *The 16PF Fifth Edition Technical Manual*. Champaign, IL: Institute for Personality and Ability Testing, Inc.
Cook, T.D., & Campbell, D.T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand McNally College Publishing.
Costa, P.T., & McCrae, R.R. (1992a). *NEO PI-R Professional Manual*. Odessa, FL: Psychological Assessment Resources.
Costa, P.T., & McCrae, R.R. (1992b). Normal personality assessment in clinical practice: The NEO personality inventory. *Psychological Assessment,, 4*, 5 – 13.
Costa, P.T., & McCrae, R.R. (1995). Domain and facets: Hierarchical personality assessment using the revised NEO personality inventory. *Journal of Personality Assessment, 64*, 21-50.
Costa, P.T., Terracciano, A., & McCrae, R.R. (2001). Gender differences in personality traits across culture: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*(2), 322-331.
Creswell, J.W. (2003). *Research Design Qualitative, Quantitative and Mixed Method Approaches*. Thousands Oak, California: SAGE Publications.
Crombie, I.K. (1996). *The Pocket Guide to Appraisal*. London: BMJ Books.
Cunha, A.D.D., & Greathead, D. (2007). Does personality matter? An analysis of code-review ability. *Communications of the ACM, 50*(5), 109-112.
Dattalo, P. (2009). A review of software for sample size determination. *Evaluation and The Health Professions, 32*(3), 229-248.
Davito, A. (1985). A review of the Myers-Briggs Type Indicator. In J. Mitchell (Ed.), *Ninth Mental Measurement Yearbook*. Lincoln: University of Nebraska Press.
De Raad, B., & Schouwenburg, H.C. (1996). Personality in learning and education: A review. *European Journal of Personality, 10*, 303-336.
DeClue, T.H. (2003). Pair programming and pair trading: Effects on learning and motivation in a CS2 courses. *Journal of Computing Sciences in Colleges, 18*(5), 49-56.
Dick, A., & Zarnett, B. (2002). Paired programming and personality traits. *Proceedings of the XP2002*, 82-85.
Digman, J.M. (1990). Personality structure: Emergence of the Five-Factor Model. *Annual Reviews Psychology, 41*, 417-440.
Dollinger, S.J., & Orf, L.A. (1991). Personality and performance in "personality": Conscientiousness and openness. *Journal of Research in Personality, 25*, 276-284.
Donaldson, S.I., & Grant-Vallone, E.J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology, 17*(2), 245-260.
Driskell, J.E., Hogan, R., & Salas, E. (1987). Personality and group performance. *Rev. Personality Social Psychology, 9*, 91-113.
Driskell, J.E., Salas, E., Goodwin, F.F., & O'Shea, P.G. (2006). What makes a good team player? Personality and team effectiveness. *Group Dynamics: Theory, Research, and*

Practice, 10(4), 249-271.
Duff, A., Boyle, E., Dunleavy, K., & Ferguson, J. (2004). The relationship between personality, approach to learning and academic performance. *Personality and Individual Differences, 36*(8), 1907 - 1920.
Dyba, T., Arisholm, E., Sjoberg, D.I.L., Hannay, J.E., & Shull, F. (2007). Are two heads better than one? On the effectiveness of pair programming. *IEEE Software, 24*(6), 12-15.
Dyba, T., & Dingsoyr, T. (2008). Empirical studies of agile software development: A systematic review. *Information and Software Technology, 50*(9-10), 833-859.
Dyba, T., & Dingsoyr, T. (2008). Strength of evidence in systematic reviews in software engineering. *Proceedings of the 2nd Int'l Symp. Empirical Software Engineering & Measurement (ESEM 2008)*, 178-187.
Dyba, T., Kampenes, V.B., & Sjoberg, D.I.L. (2006). A systematic review of statistical power in software engineering experiments. *Information and Software Technology, 48*, 745 – 755.
Dyba, T., Kitchenham, B. A., & Jorgensen, M. (2005). Evidence-based software engineering for practitioners. *IEEE Software, 22*(1), 58-65.
Elsevier, B.V. (2008). Scopus overview: What is it? Retrieved from http://www.info.scopus.com/about/
English, A., Griffith, R.L., & Steelman, L.A. (2004). Team performance: The effect of team conscientiousness and task type. *Small Group Research, 35*(6), 643-665.
Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*(1), 1-11.
Eysenck, H.J., & Eysenck, S.B.G. (1975). *Manual of the Eysenck Personality Questionnaire*. London: Hodder and Stoughton.
Farsides, T., & Woodfield, R. (2003). Individual differences and undergraduate academic success: The roles of personality, intelligence, and application. *Personality and Individual Differences, 34*(7), 1225 - 1243.
Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.
Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*(3), 429-456.
Feldt, R., Angelis, L., & Samuelsson, M. (2008). *Towards individualized software engineering: Empirical studies should collect psychometrics*. Paper presented at the CHASE'08.
Fenton, N.E., & Pfleeger, S.L. (2001). *Software Metrics: A Rigorous and Practical Approach* (2 ed.). Boston, MA: PWS Publishing Company.
Fink, A. (2005). *Conducting Research Literature Reviews. From the Internet to Paper*. Thousands Oak, CA: Sage Publication, Inc.
Francis, L., Craig, C., & Robbins, M. (2008). The relationship between the Keirsey Temperament Sorter and the short-form Revised Eysenck Personality Questionnaire. *Journal of Individual Differences, 29*(2), 116-120.
Fruyt, F.D., & Mervielde, I. (1996). Personality and interests as predictors of educational streaming and achievement. *European Journal of Personality, 10*, 405-425.
Furnham, A. (1996). The big five Vs the big four: The relationship between Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personality and Individual Differences, 21*(2), 303 - 307.
Furnham, A., Chamorro-Premuzic, T., & McDougall, F. (2003). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences, 14*, 49-66.
Gallis, H., Arisholm, E., & Dyba, T. (2003). An initial framework for research on pair programming. *Proceedings of the International Symposium on Empirical Software Engineering (ISESE'03)*, 132-142.
Gevaert, H. (2007). *Pair programming unearthed.* Unpublished M.Sc. thesis, University of Manitoba, Canada.
Goldberg, L.R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five factor models. In I. Mervielde, I. J. Deary, F. D. Fruyt & F. Ostendorf (Eds.), *Personality Psychology in Europe* (Vol. 7). Tilburg,

Netherlands: Tilburg University Press.
Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84-96.
Gorla, N., & Lam, Y.W. (2004). Who should work with whom? Building effective software project teams. *Communications of the ACM, 47*(6), 79-82.
Gosling, S.D., Vazire, S., Srivastava, S., & John, O.P. (2004). Should we trust web-based studies?: A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist February/March 2004, 59*(2), 93-104.
Gow, A.J., Whiteman, M.C., Pattie, A., & Deary, I.J. (2005). Goldberg's 'IPIP' big-five factor markers: Internal consistency and concurrent validation in Scotland *Personality and Individual Differences, 39*(2), 317-329
Gravetter, F.J., & Wallnau, L.B. (2004). *Essentials of Statistics for the Behavioral Sciences* (5 ed.). Belmont, CA: Thomson Wadsworth.
Greenhalgh, T. (2000). *How to Read a Paper: The Basics of Evidence-Based Medicine*. West Sussex, UK: BMJ Books.
Griffin, B., & Hesketh, B. (2004). Why openness to experience is not a good predictor of job performance. *International Journal of Selection and Assessment, 12*(3), 243-251.
Hanks, B. (2006). Student attitudes toward pair programming. *Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE06)*, 113-117.
Hanks, B., McDowell, C., Draper, D., & Krnjajic, M. (2004). Program quality with pair programming in CS1. *SIGCSE Bulletin, 36*(3), 176-180.
Hanks, B., Wellington, C., Reichlmayr, T., & Coupal, C. (2008). Integrating agility in the CS curriculum: Practices through values. *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education 40*(1), 19-20.
Hannay, J.E., Arisholm, E., Engvik, H., & Sjoberg, D.I.K. (2010). Effects of personality on pair programming. *IEEE Transactions on Software Engineering, 36*(1), 61-80.
Hannay, J.E., Dyba, T., Arisholm, E., & Sjoberg, D.I.K. (2009). The effectiveness of pair programming: A Meta-Analysis. *Information and Software Technology, 51*, 1110-1122.
Harris, J.A. (2004). Measured intelligence, achievement, openness to experience, and creativity. *Personality and Individual Differences, 36*, 913-929.
Heiberg, S., Puus, U., Salumaa, P., & Seeba, A. (2003). Pair-programming effect on developers productivity. *Proceedings of the 4th International Conference on Extreme Programming and Agile Processes in Software Engineering (XP 2003), LNCS 2675*, 215-224.
Highsmith, J. (2002). *Agile Software Development Ecosystems*. Boston, MA.: Addison-Wesley Longman Publishing Co. Inc.
Ho, C.-w. (2004). *Examining impact of pair programming on female students* (No. TR-2004-20). Raleigh, NC: North Carolina State University.
Howard, E.V. (2006). Attitudes on using pair-programming. *Journal of Educational Technology Systems, 35*(1), 89-103.
Hsu, B.-F., Wu, W.-L., & Yeh, R.-S. (2007). Personality composition, afftective tie and knowledge sharing: A team level analysis. *Proceedings of the Portland International Center for Management of Engineering and Technology (PICMET 2007)*, 2583-2592.
Jedlitschka, A., & Pfahl, D. (2005). Reporting guidelines for controlled experiments in software engineering. *Proceedings 2005 International Symposium on Empirical Software Engineering*, 95-104.
John, O.P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and Research* (pp. 102-138). New York/London: The Guilford Press.
Johnson, J.A. (1994). Clarification of factor five with help of the AB5C model. *European Journal of Personality, 8*(4), 311-334.
Johnson, J.A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*, 103-128.
Johnson, J.A. (2008). The IPIP-NEO personality assessment tools. Retrieved from http://www.personal.psu.edu/j5j/IPIP/
Jorgensen, M., & Shepperd, M. (2007). A systematic review of software development cost estimation studies. *IEEE Transactions on Software Engineering, 33*(1), 33-53.

Juristo, N., & Moreno, A. M. (2001). *Basics of Software Engineering Experimentation*. Boston: Kluwer Academic Publishers.

Kampenes, V.B., Dyba, T., Hannay, J.E., & Sjoberg, D.I.K. (2007). A systematic review of effect size in software engineering. *Information and Software Technology, 49*, 1073-1086.

Karn, J.S., & Cowling, T. (2006). A follow up study of the effect of personality on the performance of software engineering teams. *Proceedings of the ISESE'06*, 232-241.

Kasschau, R.A. (1985). Personality Theories. In *Psychology: Exploring Behavior* (2 ed.): Pearson Prentice Hall.

Katira, N., Williams, L., & Osborne, J. (2005). Towards increasing the compatibility of student pair programmers. *27th International Conference on Software Engineering (ICSE'05)*, 625-626.

Katira, N., Williams, L., Wiebe, E., Miller, C., Balik, S., & Gehringer, E. (2004). On understanding compatibility of student pair programmers. *SIGCSE Bulletin, 36*(1), 7-11.

Keirsey, D. (1998). *Please Understand Me II*. Del Mar, CA.: Prometheus Nemesis Book.

Keirsey, D., & Bates, M. (1984). *Please Understand Me*. Del Mar, CA: Prometheus Nemesis Book.

Khan, K.S., Kunz, R., Kleijnen, J., & Antes, G. (2003). *Systematic Reviews to Support Evidence-based Medicine,* . London: The Royal Society of Medicine Press Ltd.

Kichuk, S.L., & Wiesner, W.H. (1997). The big five personality factors and team performance: Implications for selecting successful product design teams. *Journal of Engineering and Technology Management, 14*, 195-221.

Kitchenham, B.A., & Charters, S. (2007). *Procedures for performing systematic literature reviews in software engineering*. UK: Keele University and University of Durham.

Kitchenham, B.A., Dyba, T., & Jorgensen, M. (2005). Evidence-based software engineering for practitioners. *IEEE Software, 22*(1), 58 – 65.

Kitchenham, B.A., Mendes, E., & Travassos, G.H. (2007). Cross versus within-company cost estimation studies: A systematic review. *IEEE Transactions on Software Engineering, 33*(5), 316 - 329

Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K., et al. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering, 28*(8), 721 - 734

Komarraju, M., Karau, S.J., & Schmeck, R.R. (2009). Role of the big five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences, 19*, 47-52.

Lan, L., & Lian, Z. (2010). Application of statistical power analysis - How to determine the right sample size in human health, comfort, and productivity research. *Building and Environment, 45*, 1202-1213.

Layman, L. (2006). Changing students' perceptions: an analysis of the supplementary benefits of collaborative software development. *Proceedings of the 19th Conference on Software Engineering Education & Training (CSEET'06)*, 159 - 166

Leech, N.L., Barrett, K.C., & Morgan, G.A. (2005). *SPSS for Intermediate Statistics: Use and Interpretation* (2 ed.): Mahwah, N.J. Lawrence Erlbaum Associates.

Leedy, P.D., & Ormrod, J.E. (2005). *Practical Research Planning and Design* (8th ed.): Pearson Merril Prentice Hall.

LePine, J.A. (2003). Team adaptation and postchange performance: Effects of team composition in terms of members' cognitive ability and personality. *Journal of Applied Psychology, 88*(1), 27-39.

LePine, J.A., Colquitt, J.A., & Erez, A. (2000). Adaptability to changing task contexts: Effects of general cognitive, conscientiousness, and openness to experience. *Personnel Psychology, 53*, 563-593.

Levine, J.M., & Moreland, R.L. (1990). Progress in small group research. *Annual Review of Psychology, 41*, 585-634.

Liddel, W.W., & J.W. Slocum. (1976). The effects of individual-role compatibility upon group performance: An extension of Schutz's FIRO theory. *The Academy of Management Journal, 19*(3), 413-426.

Livermore, J.A. (2006). What elements of XP are being adopted by industry practitioners? *Proceedings of the IEEE SoutheastCon 2006*, 149 – 152.

Lou, Y., Abrami, P.C., & d'Apollonia, S. (2001). Small group and individual learning with

technology: A Meta-Analysis. *Review of Educational Research, 71*(3), 449-521.
Lounsbury, J.W., Sundstrom, E., Loveland, J.M., & Gibson, L.W. (2003). Intelligence, "Big Five" personality traits, and work drive as predictors of course grade. *Personality and Individual Differences, 35*, 1231-1239.
Madeyski, L. (2006a). The impact of pair programming and test-driven development on package dependencies in object-oriented design - an experiment. *Proceedings of the 7th International Conference Product-Focused Software Process Improvement (PROFES 2006), LNCS 4034*, 278-289.
Madeyski, L. (2006b). Is external code quality correlated with programming experience or feelgood factor? *Proceeding of the7th International Conference on Extreme Programming and Agile Processes in Software Engineering (XP 2006), LNCS 4044, Springer-Verlag*, 65-74.
Matzler, K., Renzl, B., Muller, J., Herting, S., & Mooradian, T.A. (2008). Personality traits and knowledge sharing. *Journal Economic Psychology, 29*, 301-313.
McConnel, S. (1993). *Code Complete: a Practical Handbook of Software Construction*. Redmond, Washington: Microsoft Press.
McCrae, R.R., & Costa, P.T. (1989). Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *Journal of Personality, 57*(1), 17-40
McCrae, R.R., & Costa, P.T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509-516.
McCrae, R.R., & John, O.P. (1992). An introduction to the five-factor model and its application. *Journal of Personality, 60*(2), 175-215.
McDowell, C., Werner, L., Bullock, H.E., & Fernald, J. (2003). The impact of pair programming on student performance, perception and persistence. *Proceedings of the 25th International Conference on Software Engineering (ICSE'03)*, 602-607.
Mendes, E., Al-Fakhri, L., & Luxton-Reilly, A. (2005). Investigating pair-programming in a 2nd-year software development and design computer science course. *SIGCSE Bulletin, 37*(3), 296-300.
Mendes, E., Al-Fakhri, L., & Luxton-Reilly, A. (2006). A replicated experiment of pair-programming in a 2nd-year software development and design computer science course. *Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE06)*, 108-112.
Mentz, E., Walt, J.L. v. d., & Goosen, L. (2008). The effect of incorporating cooperative learning principles in pair programming for student teachers. *Computer Science Education, 18*(4), 247-260.
Miller, J., Daly, J., Wood, M., Roper, M., & Brooks, A. (1997). Statistical power and its subcomponent - missing and misunderstood concepts in empirical software engineering research. *Information and Software Technology, 39*, 285-295.
Mohammed, S., & Angell, L.C. (2003). Personality heterogeneity in teams: Which differences make a difference for team performance? *Small Group Research, 34*(6), 651 - 677.
Mohammed, S., Mathieu, J.E., & Bartlett, A.L.B. (2002). Technical-administrative task performance, leadership task performance, and contextual performance: Considering the influence of team-and task-related composition variables. *Journal of Organizational Behavior, 23*, 795-814.
Morgan, G.A., Leech, N.L., Gloeckner, G.W., & Barrett, K.C. (2004). *SPSS for Introductory Statistics. Use and Interpretation* (2nd ed.). New Jersey: Lawrence Erlbaum Associates, Inc.
Muller, M.M. (2005). Two controlled experiments concerning the comparison of pair programming to peer review. *Journal of Systems and Software, 78*(2), 166-179.
Muller, M.M. (2006). A preliminary study on the impact of a pair design phase on pair programming and solo programming. *Information and Software Technology, 48*(5), 335-344.
Muller, M.M., & Padberg, F. (2004). An empirical study about the feelgood factor in pair programming. *Proceedings of the 10th International Symposium on Software Metrics (METRIC 2004)*, 151-158.
Murphy, K.R., & Myors, B. (2003). *Statistical Power Analysis: A simple and General Model for Traditional and Modern Hypothesis Tests* (2 ed.). New Jersey: Lawrence Erlbaum Associates.
Murray, J.B. (1990). Review of research on the Myers-Briggs Type Indicator. *Perceptual and*

Motor Skills, 70, 1187-1202.
Myers, I.B., McCaulley, M.H., Quenk, N.L., & Hammer, A. (1998). *MBTI Manual (A guide to the development and use of the Myers Briggs type indicator)* (3rd ed.): Consulting Psychologists Press.
Myers, I.B., & Myers, P.B. (1995). *Gifts differing: Understanding Personality Type*. Mountain View, CA: Davies-Black Publishing.
Myers, J.L., & Well, A.D. (2003). *Research Design and Statistical Analysis* (2 ed.). New Jersey: Lawrence Erlbaum Associates, Inc.
Nagappan, N., Williams, L., Ferzli, M., Wiebe, E., Yang, K., Miller, C., et al. (2003). Improving the CS1 experience with pair programming. *SIGCSE Bulletin, 35*(1), 359-362.
Nawrocki, J., & Wojciechowski, A. (2001). Experimental evaluation of pair programming. *in Proceedings European Software Control and Metrics (ESCOM)*, 269-276.
Neuman, G.A., Wagner, S.H., & Christiansen, N.D. (1999). The relationship between work-team personality composition and the job performance of teams. *Group & Organization Management, 24*(1), 28 - 45.
Nguyen, N.T., Allen, L.C., & Fraccastoro, K. (2005). Personality predicts academic performance: Exploring the moderating role of gender. *Journal of Higher Education Policy and Management, 27*(1), 105 - 116.
Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Science Education*(Online First).
Nosek, J.T. (1998). The case for collaborative programming. *Communications of the ACM, 41*(3), 105-108.
O'Connor, M.C., & Paunonen, S.V. (2007). Big five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43*, 971-990.
Ogot, M., & Okudan, G.E. (2006). The five-factor model personality assessment for improved student design team performance. *European Journal of Personality, 31*(5), 517-529.
Pallant, J.F. (2007). *SPSS Survival Manual: A step by step guide to data analysis using SPSS for Windows (Version 15)* (3rd ed.). Crows Nest, N.S.W.: Allen & Unwin.
Paunonen, S.V., & Ashton, M.C. (2001). Big five predictors of academic achievement. *Journal of Research in Personality, 35*, 78-90.
Peeters, M.A.G., Tuijil, H.F.J.M.V., Rutte, C.G., & Reymen, I.M.M.J. (2006). Personality and team performance: A Meta-Analysis. *European Journal of Personality, 20*, 377-396.
Peslak, A.R. (2006). The impact of personality on information technology team projects. *Proceedings of the SIGMIS-CPR'06*, 273 - 279.
Petticrew, M., & Roberts, H. (2006). *Systematic Reviews in the Social Sciences: A Practical Guide*: Blackwell Publishing.
Pfleeger, S.L. (1995). Experimental design and analysis in software engineering. *Annals of Software Engineering, 1*(1), 219-253.
Phillips, P., Abraham, C., & Bond, R. (2003). Personality, cognition, and university students' examination performance:. *European Journal of Personality, 17*, 435-448.
Phongpaibul, M., & Boehm, B. (2006). An empirical comparison between pair development and software inspection in Thailand. *Proceedings of the 5th ACM-IEEE International Symposium on Empirical Software Engineering (ISESE'06)*, 85 - 94
Pickard, L.M., Kitchenham, B.A.J., & Jones, P.W. (1998). Combining empirical results in software engineering. *Information and Software Technology, 40*(4), 811-821.
Pieterse, V., & Kourie, D.G. (2006). Software engineering team diversity and performance. *Proceedings of the South African Institute for Computer Scientists and Information Technologists (SAICSIT), 2006*, 180-186.
Pittenger, D.J. (1993). Measuring the MBTI...and coming up short. *Journal of Career Planning and Employment*, 48 - 53.
Poropat, A.E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338.
Pulford, B.D., & Sohal, H. (2006). The influence of personality on higher education students' confidence in their academic abilities. *Personality and Individual Differences, 41*(8), 1409 - 1419.
Puus, U., Seeba, A., Salumaa, P., & Heiberg, S. (2004). Analyzing pair-programmer's satisfaction with the method, the result, and the partner. *Proceedings of the 5th International Conference on Extreme Programming and Agile Processes in Software Engineering (XP 2004), LNCS 3092*, 246-249.

Rosenthal, R., & DiMatteo, M.R. (2001). Meta-analysis: recent development in quantitative methods for literature review. *Annual Review of Psychology, 52*, 59-82.

Rostaher, M., & Hericko, M. (2002). Tracking test first pair programming - An experiment. *Proceedings of the XP/Agile Universe 2002, LNCS 2418*, 174-184.

Rutherfoord, R.H. (2001). Using personality inventories to help form teams for software engineering class projects. *Proceedings of the ITiCSE 2001*, 73-76.

Salleh, N., Mendes, E., & Grundy, J. (2010). Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review. *IEEE Transactions on Software Engineering (to appear) DOI: 10.1109/TSE.2010.59*

Salleh, N., Mendes, E., Grundy, J., & Burch, G.S.J. (2009). An empirical study of the effects of personality in pair programming using the five-factor model. *Proceedings of the 3rd ACM-IEEE Int'l Symposium on Empirical Software Engineering & Measurement (ESEM 2009)*, 214-225.

Salleh, N., Mendes, E., Grundy, J., & Burch, G.S.J. (2010a). An empirical study of the effects of conscientiousness in pair programming using the five-factor personality model. *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering (ICSE 2010), 1*, 577-586.

Salleh, N., Mendes, E., Grundy, J., & Burch, G.S.J. (2010b). The effects of neuroticism on pair programming: An empirical study in the higher education context. *Proceedings of the 4th ACM-IEEE Int'l Symposium on Empirical Software Engineering and Measurement (ESEM 2010)*.

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar big-five markers. *Journal of Personality Assessment, 63*, 506-516.

Schloss, P.J., & Smith, M.A. (1999). *Conducting Research*. Upper Saddle River, NJ: Prentice Hall.

Schmitt, N. (2008). The interaction of neuroticism and gender and its impact on self-efficacy and performance. *Human Performance, 21*, 49-61.

Schriesheim, C., Hinkin, T., & Podsakoff, P. (1991). Can ipsative and single-item measures produce erroneous results in field studies of French & Raven's (1959) five bases of power? *Journal of Applied Psychology, 76*, 106-144.

Sfetsos, P., Stamelos, I., Angelis, L., & Deligiannis, I. (2006). Investigating the impact of personality types on communication and collaboration-viability in pair programming - an empirical study. *Proceedings of the 7th International Conference on Extreme Programming and Agile Processes in Software Engineering (XP 2006), LNCS 4044*, 43-52.

Sfetsos, P., Stamelos, I., Angelis, L., & Deligiannis, I. (2009). An experimental investigation of personality types impact on pair effectiveness in pair programming. *Empirical Software Engineering, 14*, 187-226.

Sison, R. (2008). Investigating pair programming in a software engineering course in an Asian setting. *Proceedings of the 15th Asia-Pacific Software Engineering Conference (APSEC)*, 325-331.

Sison, R. (2009). Investigating the effect of pair programming and software size on software quality and programmer productivity. *Proceedings of the 16th Asia-Pacific Software Engineering Conference (APSEC)*, 187-193.

Sjoberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanovic, A., Liborg, N.-K., et al. (2005). A survey of controlled experiments in software engineerings. *IEEE Transactions on Software Engineering, 31*(9), 733.

Slavin, R. E. (1980). Cooperative learning. *Review of Educational Research, 50*(2), 315-342.

Slavin, R. E. (1990). *Cooperative Learning: Theory, Research and Practice*. Englewood Cliffs, N.J.: Prentice Hall.

Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). *Quality in qualitative evaluation: A framework for assessing research evidence*. London: Government Chief Social Researcher's Office.

Staples, M., & Niazi, M. (2007). Experiences using systematic review guidelines,. *Journal of Systems and Software, 80*(9), 1425-1437.

Staudinger, U.M., Maciel, A.G., Smith, J., & Baltes, P.B. (1998). What predicts wisdom-related performance? A first look at personality, intelligence, and facilitative experiential contexts. *European Journal of Personality, 12*, 1-17.

Steiner, L.D. (1972). *Group Process and Productivity*. New York and London: Academic Press.

Stevens, J.P. (2002). *Applied Multivariate Statistics for the Social Sciences* (4 ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Tabachnick, B.G., & Fidell., L.S. (2001). *Using Multivariate Statistics* (4 ed.). Boston, MA: Allyn and Bacon.

Thomas, L., Ratcliffe, M., & Robertson, A. (2003). Code warriors and code-a-phobes: A study in attitude and pair programming. *SIGCSE Bulletin, 35*(1), 363-367.

Trochim, W.M.K. (2006). Research Methods Knowledge Base. 2nd Edition. from http://www.socialresearchmethods.net/kb/considea.php (version current as of October 20, 2006).

VanDeGrift, T. (2004). Coupling pair programming and writing: learning about students' perceptions and processes. *ACM SIGCSE Bulletin, 36*(1), 2-6.

Vanhanen, J., & Lassenius, C. (2005). Effects of pair programming at the development team level: an experiment. *Proceedings of the 2005 International Symposium on Empirical Software Engineering (ISESE 05), IEEE CS Press*, 336 - 345.

Vianen, A.E.M. v., & Dreu, C.K.W.D. (2001). Personality in teams: Its relationship to social cohesion, task cohesion, and team performance. *European Journal of Work and Organizational Psychology, 10*(2), 97-120.

Vygotsky, L.S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: MIT Press.

Walle, T., & Hannay, J.E. (2009). Personality and the nature of collaboration in pair programming. *Proceedings of the 3rd Int'l Symp. Empirical Software Engineering & Measurement (ESEM 2009)*, 203-213.

Weinberg, G.M. (1971). *The Psychology of Computer Programming*. New York, USA: Van Nostrand Reinhold.

Werner, L.L., Hanks, B., & McDowel, C. (2004). Pair-programming helps female computer science students. *Journal of Educational Resources in Computing (JERIC), 4*(1).

Whyte, J.J. (2006). The Use of Surrogate Outcome Measures. A Case Study: Home Prothrombin Monitors. In K. M. Becker & J. J. Whyte (Eds.), *Clinical Evaluation of Medical Devices: Principles and Case Studies* (2 ed.). New Jersey: Humana Press.

Wikipedia (2010). Eysenck Personality Questionnaire. Retrieved 2 June 2010 23:58 UTC from http://en.wikipedia.org/wiki/Eysenck_Personality_Questionnaire

Williams, L. (2000). *The Collaborative Software Process.* Unpublished PhD Dissertation, University of Utah, Utah.

Williams, L., & Kessler, R.R. (2002). *Pair Programming Iluminated*. Boston, MA, USA: Addision-Wesley Longman Publishing Co., Inc.

Williams, L., Kessler, R.R., Cunningham, W., & Jeffries, R. (2000). Strengthening the case for pair programming. *IEEE Software, 17*(4), 19-25.

Williams, L., Layman, L., Osborne, J., & Katira, N. (2006). Examining the compatibility of student pair programmers. *Proceedings of the Conference on AGILE 2006 (AGILE'06), IEEE Computer Society*, 411-420.

Williams, L., McDowell, C., Nagappan, N., Fernald, J., & Werner, L. (2003). Building pair programming knowledge through a family of experiments. *Proceedings 2003 International Symposium on Empirical Software Engineering (ISESE 2003)*, 143-152.

Williams, L., Yang, K., Wiebe, E., Ferzli, M., & Miller, C. (2002). Pair programming in an introductory computer science course: Initial results and recommendations. *Proceedings 17th Annual ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPLSA 2002)*.

Williams, L.A., & Kessler, R.R. (2000). The effects of "pair-pressure" and "pair-learning" on software engineering education. *13th Conference on Software Engineering Education and Training, IEEE Computer Society*, 59-65.

Witt, L.A., Barrick, M.R., Burke, L.A., & Mount, M.K. (2002). The interactive effects of conscientiousness and agreeableness on job performance. *Journal of Applied Psychology, 87*(1), 164-169.

Wohlin, C., Runeson, P., Host, M., Ohlsson, M.C., Regnell, B., & Wesslen, A. (2000). *Experimentation in Software Engineering: An Introduction*. Boston: Kluwer Academic Publisher.

Xu, S., & Rajlich, V. (2006). Empirical validation of test-driven pair programming in game development. *5th IEEE/ACIS International Conference on Computer Information Science - In Conjunction with 1st IEEE/ ACIS Workshop on Component-Based Software Engineer Architecture and Reuse*, 500-505.

Yuan, K., & Maxwell, S. (2005). On the Post Hoc Power in Testing Mean Differences. *Journal of Educational and Behavioral Statistics, 30*(2), 141-167.

Zemke, R. (1992). Second thoughts about the MBTI. *Training, 29*(4), 43-47.

Zyphur, M.J., Bradley, J.C., Landis, R.S., & Thoresen, C.J. (2008). The effects of cognitive ability and conscientiousness on performance over time: A censored latent growth model. *Human Performance, 21*, 1-27.