Empirical Software Engineering, DOI 10.1007/s10664-012-9238-4, © Springer 2013

Investigating the effects of personality traits on pair programming in a higher education setting through a family of experiments

N. Salleh[1], E. Mendes[2], J. Grundy[3]

[1]*Department of Computer Science, International Islamic University Malaysia*
P.O. Box 10, 50728 Kuala Lumpur, Malaysia
norsaremah@iium.edu.my

[2]*School of Computing, Blekinge Istitute of Technology*
37179, Karlskrona, Sweden
emilia.mendes@bth.se

[3]*Faculty of Information and Communication Technology, Swinburne University of Technology, P.O. Box 218, Hawthorn, Victoria, Australia*
jgrundy@swin.edu.au

Abstract

Evidence from our systematic literature review revealed numerous inconsistencies in findings from the Pair Programming (PP) literature regarding the effects of personality on PP's effectiveness as a pedagogical tool. In particular: i) the effect of differing personality traits of pairs on the successful implementation of pair-programming (PP) within a higher education setting is still unclear, and ii) the personality instrument most often used had been Myers-Briggs Type Indicator (MBTI), despite being an indicator criticized by personality psychologists as unreliable in measuring an individual's personality traits. These issues motivated the research described in this paper. We conducted a series of five formal experiments (one of which was a replicated experiment), between 2009 and 2010, at the University of Auckland, to investigate the effects of personality composition on PP's effectiveness. Each experiment looked at a particular personality trait of the Five-Factor personality framework. This framework comprises five broad traits (Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), and our experiments focused on three of these - Conscientiousness, Neuroticism, and Openness. A total of 594 undergraduate students participated as subjects. Overall, our findings for all five experiments, including the replication, showed that Conscientiousness and Neuroticism did not present a statistically significant effect upon paired students' academic performance. However, Openness played a significant role in differentiating paired students' academic performance. Participants' survey 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.

Keywords: Pair programming, formal experiment, personality traits, five-factor model, higher education

1. Introduction

The pair programming (PP) technique has been in use in academic education and industry domains for over ten years. It involves teams of two people sitting side by side, using only one computer, and working collaboratively on the same design, algorithm, code or test. One is normally the "driver", who is responsible for typing the code and has control over the resources (i.e. computer, mouse and keyboard). The partner is usually known as the "navigator" or "observer", and takes the responsibility for observing how the driver works in order to detect errors and offer ideas in solving a problem (Beck, 1999). In an academic context, much research on PP has been conducted to determine benefits of the technique and to understand how the practice could help improve students' learning outcomes. For example, research evidence suggests PP can enhance students' enjoyment, increase students' confidence level, improve course completion rates, and facilitate students in working more efficiently on programming tasks (McDowell et al., 2003; Williams et al., 2003).

However, despite these benefits, PP was reported to be problematic most often when pairs are "incompatible" (Layman, 2006). Previous studies reported that students who experience PP with an incompatible partner disliked the collaborative work (Layman, 2006; Thomas et al., 2003). Such discomfort or incompatibility of 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. Therefore, the selection of personality traits as a variable may provide an advantage in overcoming the problem of bad pairing experiences reported in some PP studies (Layman, 2006; Ho, 2004). The findings from our systematic literature review (SLR) indicate that students prefer to work with a partner who is at a similar skill level as their own (Salleh et al., 2011). Cockburn & Williams (2001) highlight that understanding the social aspects of PP is critical for attaining the success with 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 behaviors and opinions, understanding how two students can best work together is imperative to the success of PP as a pedagogical tool (Choi, 2004; Sfetsos et al., 2009).

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 pair compatibility can be potentially affected by human-related factors such as personality (Sfetsos et al., 2006). Existing literature in Agile methods suggests that developers' personality is one of PP's most critical success factors (Cockburn, 2002). Weinberg (1971) also asserted that a programmer's personality is an important parameter for determining if the programmers' job is a success. He 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). Therefore, understanding how personality affects or relates to PP's effectiveness is an important aspect of using the technique.

The aim of our research was to maximize the effectiveness of PP as a pedagogical tool for use in higher education through understanding the impact that the variation in the personality composition[1] of paired students has towards their academic performance. Thus, our overarching research question was: "*Do personality traits affect academic performance of undergraduate students practicing PP?*" Our approach to answering this was to conduct a series of four formal experiments (*Exp 1* [Salleh et al., 2009], *Exp 2* [Salleh et al., 2010a], *Exp 3* [Salleh et al., 2010b], *Exp 4* [Salleh et al., 2011a]), and a replicated experiment (*Exp 5*), all targeted at Computer Science (CS) undergraduate courses at the University of Auckland. We focused our investigation on the three major traits from the Five Factor Model (FFM[2]): Conscientiousness (*Exp 1*, *Exp 2*, *Exp 5*), Neuroticism (*Exp 3*), and Openness to experience (*Exp 4*). These traits were chosen because evidence shows that they are important and relevant for influencing academic success in tertiary education (De Raad & Schouwenburg, 1996). We also aimed to investigate a second overarching research question: "*Would personality trait composition in PP affect the level of satisfaction and confidence on students when pairing?*"

The contribution of the research described herein is fourfold:

- extending our data analysis for all formal experiments (*Exp 1 – Exp 4*) with a power analysis, strengthening results' discussion and interpretation.

[1] Our focus was on between-pair differences.
[2] FFM is detailed in Section 3

- combining results from the four experiments (*Exp 1 – Exp 4*) to enable us to answer our overarching research questions. Each experiment investigated a single personality trait and its effect on PP's effectiveness, using programming tasks as the experimental object. However, it is paramount to consider all of the results together so that we can better understand the role personality may have on PP's effectiveness as a pedagogical tool.
- a replicated formal experiment (*Exp 5*), replicating *Exp 2*, where the effect of personality trait Conscientiousness on PP's effectiveness was investigated using software design tasks instead of programming.
- combining results from *Exp 2* and *Exp 5* in order to investigate whether trends were similar despite differences in tasks.

The results of this research can be used to better inform teachers about the implications of team personalities on academic performance when employing PP. Specifically, their team formation approaches can be influenced to maximize successful learning outcomes. We believe our research is also useful to guide future research into PP team composition based on personality traits.

Section 2 presents an overview of related work and motivation for this research. Section 3 briefly describes the five-factor personality model. Sections 4 and 5 describe the experimental definition and the four formal experiments. Section 6 presents the aggregated results from combining those four experiments and a detailed discussion of our findings. Section 7 details the replicated study (*Exp 5*), and compares its findings to those obtained in *Exp 2*. Finally, Section 8 concludes our work and summarizes key future directions for research.

2. Motivation & Related work

The issue of personality in PP has been addressed in a number of studies conducted both in academic and industrial settings, where the central theme has focused on investigating the impact of personality on the performance of teams and individuals practicing PP (e.g. Williams et al., 2006, Choi et al., 2008, Sfetsos et al., 2006, Hannay et al., 2010). Based on the results from our systematic literature review (SLR) of PP in higher education, we found evidence that only 23% of the included studies had empirically investigated factors that may affect PP's success, one of them being personality (Salleh et al., 2011). Empirical

evidence from our SLR suggested that personality was one of the most common factors investigated in previous PP studies, which implies that personality is intrinsically related to PP's success.

In assessing personality, the Myers-Briggs Type Indicator (MBTI) has been used as a personality measure in most existing PP studies in academic settings (Salleh et al., 2011). Others have used the Keirsey Temperament Sorter (KTS) (Sfetsos et al., 2006) and most recently some studies have applied the big-five or Five-Factor personality model (Hannay et al., 2010), including our own (Salleh et al., 2009, Salleh et al., 2010a, Salleh et al., 2010b, Salleh et al., 2011a). The MBTI is one of the most widely-used personality assessment tools employed to measure an individual's personality preferences. It has been commonly used in the area of training and consultancy (Furnham, 1996) and also widely-used by researchers in the Information Systems (IS) and Software Engineering (SE) 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 et al., 1991; McCrae & Costa, 1989). The MBTI instrument, which uses as its basis the psychodynamic type theory of Jungian concept, has been 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).

The findings reported in five PP studies that investigated personality using the MBTI were quite diverse and thus inconclusive on whether personality could significantly affect pair programmers' productivity (Salleh et al., 2011). 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; Layman, 2006; Williams et al., 2006). Different outcomes reported from these studies could be accounted for by the differential set of instruments and personality frameworks used to measure personality and variation in study context, making it difficult to generalize results.

Due to 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[3]. The FFM was chosen in our 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 (Burch & Anderson, 2008; Barrick, Mount, & Judge, 2001). There is a growing consensus among personality trait researchers that FFM consists of a robust taxonomy of personality (Neuman et al., 1999; Farsides & Woodfield, 2003; Burch & Anderson, 2008). In terms of its validity and reliability, FFM is generally accepted by psychologists who suggest that its traits adequately represent human personality attributes (Barrick & Mount, 1991; Barrick et al., 1998). Relevant findings from industry usage are referred to as there are not many studies in an academic setting within the IS/IT/SE domain that have used the FFM.

To date, empirical findings using the FFM reported low support for the effects of personality in PP. Hannay et al. (2010) showed personality as being only a moderate predictor for performance. They suggested that the performance of pair programmers may also be affected by other factors such as expertise, and task complexity. Another 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 the personality factor Extraversion is positively correlated with software quality, and teams with higher aggregate on Agreeableness and Conscientiousness achieved the highest job satisfaction.

In the IS and SE literature, 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 more effective in terms of the team's performance (Pieterse & Kourie, 2006). Rutherfoord (2001)

[3] See detail about FFM in Section 3

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) reported 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 personality to be the strongest predictor compared with cognitive ability. When it comes to performing code-review tasks, Cunha and Greathead (2007) reported that people who were more intuitive performed better than those who were less intuitive.

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.

Team composition refers to the process of arranging a team based on its members' attributes such as personality, expertise, 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 (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. Most studies in the computing domain (i.e. IS/CS/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 higher team performance (e.g. Peeters et al., 2006; Kichuk & Wiesner, 1997). For instance, a team consisting of conscientious members is reported to perform better compared with a team with a heterogeneous level of Conscientiousness (Peeters et al., 2006).

In studies that measured academic performance using FFM, there was 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. Therefore there is a lack of evidence on causal-effect type studies in the personality-related literature. Boekaerts (1996) mentions that a 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 study the effects of personality traits on various outcome variables such as achievement and learning strategies.

3. The Five-Factor Model

The Five-Factor Model (FFM), also known as the "Big Five", is a taxonomy of personality traits (see Figure 1). It comprises five broad personality traits that together provide a structure for categorizing dimensions of differences in human personality (McCrae & John, 1992).The five traits were 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 & 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).

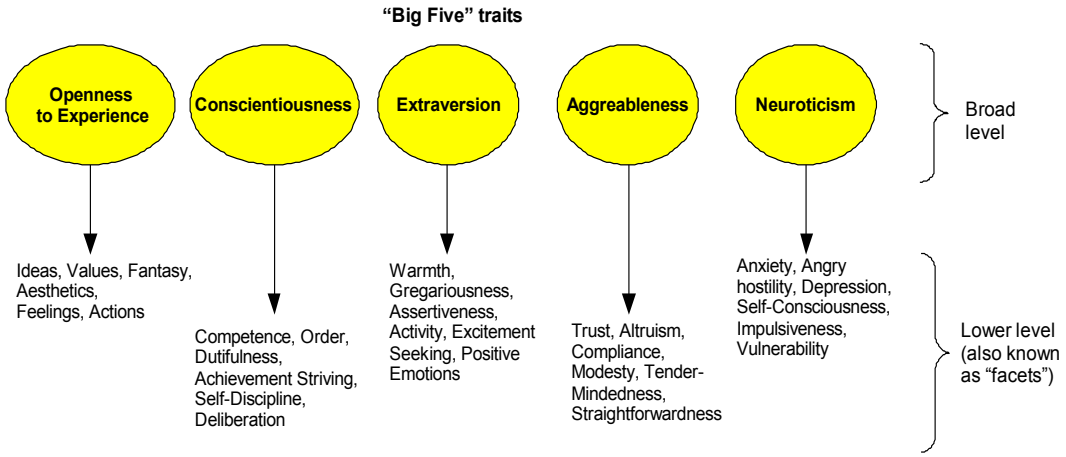

Figure 1 The Five-Factor Model (FFM)

Although the FFM and the Big Five have different theoretical underpinnings - the former is associated with the emergence of personality factors based on questionnaires, the later based on a lexical hypothesis approach – they are both referring to the same dimension of personality factors. The Big Five represents a global term for personality models that consists of five factors. The FFM is a specific type of the Big Five that consists of the following traits:

- ***Conscientiousness*** is concerned with one's achievement orientation. Those who have a high score tend to be hardworking, organized, able to complete tasks thoroughly and on-time, and reliable. On the other hand, low Conscientiousness relates to negative traits such as being irresponsible, impulsive, and disordered (Driskell et al., 2006)
- ***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.
- ***Agreeableness*** refers to positive traits such as cooperativeness, kindness, trust, and warmth. A person who is low on Agreeableness tends to be skeptical, selfish, and hostile. A team that requires a high level of collaboration or cooperation can benefit from agreeable team members.
- ***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). Neuroticism is also reported to be consistently related to self-efficacy (Schmitt, 2008).

- ***Openness to experience*** describes intellectual, cultural, or creative interest (Driskell et al., 2006). Someone who is high in Openness tends to appear as imaginative, broad-minded, 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).

According to John and Srivastava (1999), the five personality 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 1 shows the 30 facet scales of NEO-PI-R's inventory, identified and empirically validated by Costa & McCrae (Costa & McCrae, 1995).

In terms of operationalizing the FFM, there are various instruments developed to measure personality using the FFM traits. *NEO Five-Factor Inventory* (NEO-FFI) is one of the instruments that is well accepted, widely assessed, and extensively used to measure the 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. Later, a Web-based instrument known as *International Personality Item Pool* (IPIP) was developed by Goldberg and Johnson (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, improving reliability of the inventory. Such an automated instrument is much more efficient compared with any paper-based personality instruments (Goldberg et al., 2006).

In terms of the validity of the scales used to measure personality, the IPIP-NEO is reported to have good reliability compared to other established personality instruments (e.g. NEO-PI-R) (Johnson, 2005; Goldberg, 2006). 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 0.7 (Morgan et al., 2004; Pallant, 2007).

In comparison with the MBTI, the FFM was derived based on factor-analytic studies, whereas the MBTI was developed based on Jung's theory of psychological types (Furnham, 1996). The MBTI categorizes individual behavior into four dimensions of personality type: *Extraversion (E)* vs *Introversion* (I), *Sensing* (S) vs *Intuition* (N), *Thinking* (T) vs *Feeling* (F), and *Judging* (J) vs *Perceiving* (P) (Myers et al., 1998). Thus, 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 (e.g. ENFJ). In the 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, the 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) compared with non-parametric statistical tests that need to be employed with other frameworks (Feldt et al., 2008).

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, 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*)*; and *Rational* (seeking mastery and self control, typically intuitive, practical and realistic).

The major difference between 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 et al., 2008). In

classifying personality type, MBTI emphasizes the extraversion and introversion (i.e. expressive/attentive) dichotomy, while KTS stresses importance of the sensing/intuition (i.e. concrete/abstract) perspective (Francis et al., 2008).

4. Research Methodology

4.1 Research Objectives

Our research objectives are outlined using the Goal/Question/Metric (GQM) framework defined by Basili et al., 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 1, and the purpose of all the four 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 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 performance (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.

4.2 Research 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; Karn and Cowling, 2006; Busato et al., 2000; and Pieterse & Kourie, 2006). However, many of these

studies were conducted in the context of teams consisting of four to five members. The effects of personality were investigated in numerous PP studies involving peer or pair collaboration (Salleh et al., 2011), but most using the MBTI personality framework. As far as we are aware, there are no available theories that link the 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). In order to investigate the effect of personality differences on PP's effectiveness, the following overarching hypothesis was proposed:

H_O: Differences in personality traits do not affect the academic performance of undergraduate students who pair programmed.

which is contrasted by the following alternative hypothesis:

H_A: Differences in personality traits affect the academic performance of students who pair programmed.

In this research we focused on the three traits part of the FFM reported to be important educationally and relevant for higher education: Conscientiousness, Openness to experience and Neuroticism (De Raad & Schouwenburg, 1996). A motivation for each of these three traits will be detailed below, and followed by more specific hypotheses aimed at each particular trait.

Previous findings showed Conscientiousness to consistently positively predict educational success (Busato et al., 2000; Duff et al., 2004; Chamorro-Premuzic & Furnham, 2003b). High Conscientiousness appears to always be related to being a high achiever, being organized, and being thorough. In contrast, low Conscientiousness possesses the opposite traits such as a low need for achievement, being unprepared and being disorganized (McCrae& John, 1992). Thus, 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:

H1_O: Differences in Conscientiousness level do not affect the academic performance of students who pair programmed.

which is contrasted by the following alternative hypothesis:

H1_A: Differences in Conscientiousness level affect the academic performance of students who pair programmed.

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; Chamorro-Premuzic & Furnham, 2003b). However, there is some evidence from organizational psychology that in certain conditions anxiety and Neuroticism may actually facilitate performance (Burch & Anderson, 2008). Emotional stability is consistently related to self-efficacy, which in turn, affects performance (Schmitt, 2008; Barrick et al., 1998). 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:

H2_O: Differences in Neuroticism level do not affect the academic performance of students who pair programmed.

which is contrasted by the following alternative hypothesis:

H2_A: Differences in Neuroticism level affect the academic performance of students who pair programmed.

Personality research on team settings reports that teams composed of highly open 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 as those who scored high on this trait are more adaptable and capable of handling changes 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). We conjecture that paired students' academic performance may be influenced by the level of Openness to experience. Hence, the following hypothesis was proposed:

H3_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:

H3_A: Differences in levels Openness to experience affect the academic performance of students who pair programmed.

4.3 Context

The hypotheses detailed in Section 4.2 were investigated in a series of PP experiments conducted at the University of Auckland between the periods of 2009-2010 involving an undergraduate course: Principles of Programming (COMPSCI 101). COMPSCI 101 is an introductory course for first-year students learning an object-oriented programming language, Java. Students learn about basic programming concepts and create small applications in their assignments.

Figure 2 details the sequence and timeline in which each of the four formal experiments took place, and also provides a broader overview on the entire research process employed herein. Note that prior to conducting the experiments, we obtained approval from the University of Auckland's Human Participants Ethics Committee. As shown in Figure 2, each of the experiments we conducted investigated a different set of personality traits.

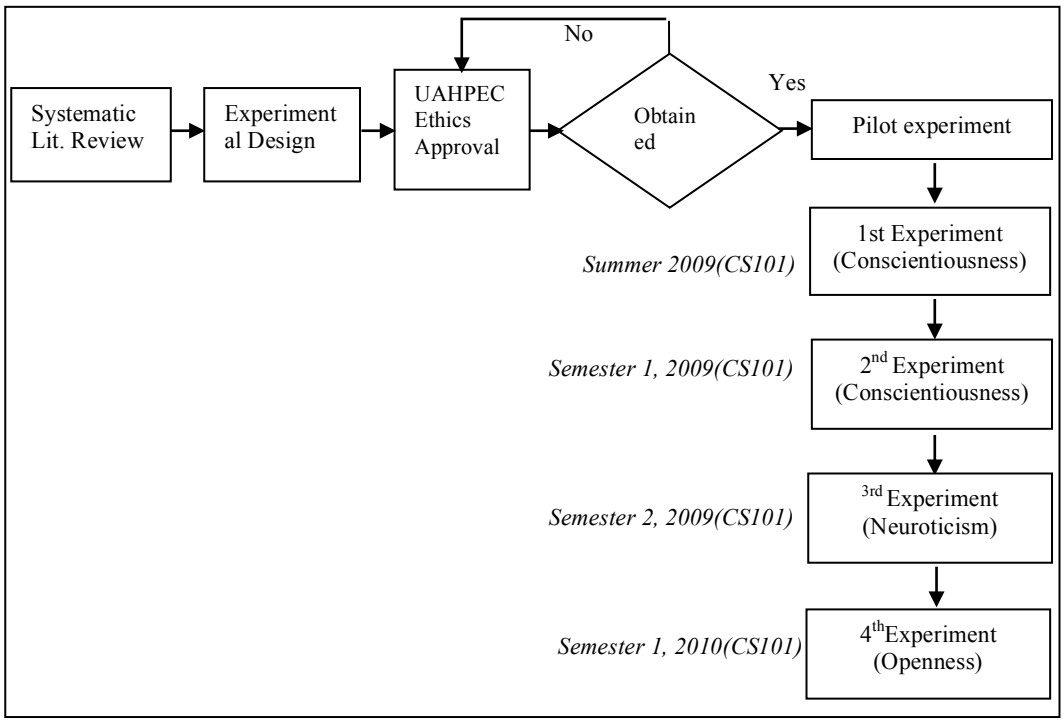

Figure 2 Research process

The sequence of the experiments was decided upon dynamically as the research progressed and also depended on the personality data distributions (which should be approximately normally distributed) derived from the sample. This was done for the purpose of complying with the type of statistical tests employed in this research. The second formal experiment on the Conscientiousness trait was carried out due to the issue of pair formation strategy used in the first experiment.

4.3 Research Design

A model of the research design used is shown in Figure3. This was derived from the initial framework for research on PP proposed by Gallis, Arisholm & Dyba (2003). It shows 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" or variable 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. Participants were allocated into pairs according to their personality level. For instance, pair configuration for 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.

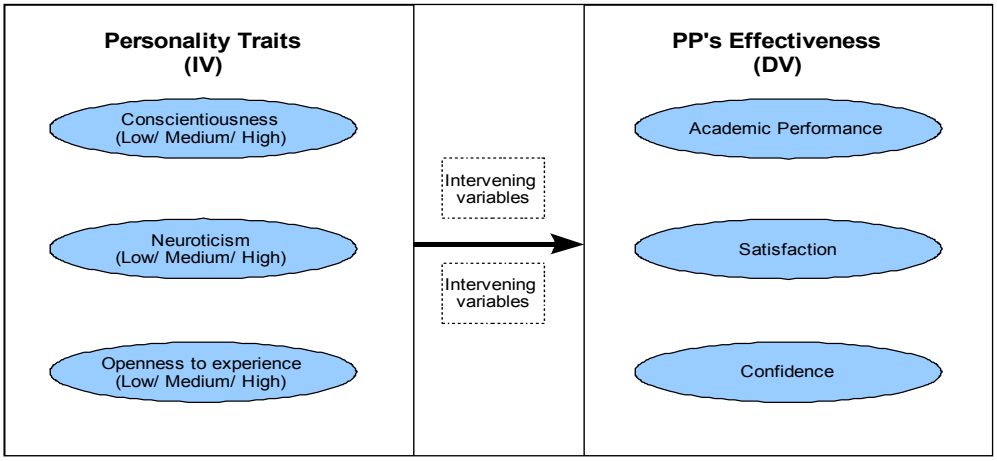

Figure3 Visual model of research design

The research design employed in all four formal experiments was a "single-factor between-group design" (Morgan et al., 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 where the groups used related to combinations of 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". This design was not used in the later experiments due to the issue of pair formation: 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).

PP's effectiveness was the outcome to be measured in 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., 2011). 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. The scores that the students obtained were considered as a proxy for measuring their learning/performance.

The tutorials' topic varied from week to week and therefore 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, tasks and exercises assigned remained the same throughout a week. In this regard, the blocking variable applied to all the experiments was the topic for exercises.

Academic performance was measured using assignments, a midterm test, and final exam scores. We used multiple-choice and fill in the blanks type of questions for tests and exams. Assigments were evaluated based on the answer scheme.

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. Personality scores were 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).

4.4 Instrumentation and Materials

There were six types of instruments 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. Participation in this research was on a voluntary basis. Therefore subjects were given the right to withdraw from the study at any time before the end of the semester. Participants who were willing to participate are given a consent form. The consent form lists the statements indicating the nature or conditions of participation and participants sign to indicate their agreement.

In order to assess the suitability and clarity of the instruments proposed (e.g. questionnaires), we conducted a pilot study prior to actual experiment and data collection. The purpose of the pilot was to validate the instruments to be used and the research design. The pilot was carried out in 2008 with the participation of 31 Computer Science students from a second-year software programming and design course at the University of Auckland. We made several amendments to the PP questionnaire in order to improve its clarity including re-phrasing questions and adding an open-ended question to allow participants to state comments.

The short version of the *International Personality Item Pool Representation of NEO PI-R*^{TM} (IPIP-NEO) was employed to measure personality traits of

participants[4]. The IPIP-NEO was developed based on the *International Personality Item Pool* (IPIP), a scientific collaboration for the development of personality measurement scale (Goldberg, 2006). The original version of IPIP-NEO contains 300 items whereas the short version contains 120 items (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 paper-based (Buchanan, Johnson & Goldberg, 2005; Gosling, Vazire, Srivastava, & John, 2004).

Each item in the IPIP-NEO personality inventory is indicated using a 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) 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: Low (lowest 30%), Medium (Middle 40%), High (highest 30%).

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. 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 A). 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[5]) software was used in order to automate the process of pairing formation. It is a Java-based application that connects to a MySQL database server and runs under the Eclipse 3.2 environment. Based on the weekly list of

[4] The personality test is available at the public domain http://www.personal.psu.edu/ /j5j/IPIP/
[5] The source code is available from the first author

students provided by the tutor, PALLOC generates a pairing list in Microsoft Excel's document format. This pairing list was used by tutors to organise teams for each tutorial activity. Students were paired with different partner for each tutorial session in order to diversify their pairing experience.

4.5 Experimental Procedure

All the experiments shared the same experimental procedure: At the start of the academic semester we approached the experimental subjects in the first lecture for the chosen course. Students were given an overview of the experiment, including a brief explanation on PP. The PIS, the demographic survey form and the consent form were distributed. The participants' personality profiles were 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 personality score of a specific 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.

All four experiments were carried out as part of COMPSCI 101 (Principles of Programming). This course is for first-year undergraduate students, and 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. Each 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. 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 et al., 2000), students worked with their partner for an initial period of 30 minutes. They were then required to swap 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, contributing towards their final

grade. 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.

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 and 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 A). The experiments aimed to measure the effect of pair personality composition towards the academic performance of the paired students. Thus, the same research design was used every week until the end of the semester.

5. Results from the four experiments

This Section describes the results of the series of experiments including the demographics relating to the sample population used. In total, 517 undergraduate students participated in our experiments (see Table 2). In the first experiment (*Exp 1*), subjects were undergraduate students age ranged between 19 to 30 years. Sixty-five (65%) of the subjects were male, 35% female. In the second experiment (*Exp 2*), 77% of the subjects were male, 23% females, and subjects' age ranged from 19 to 52 years (median = 19 years). In the third experiment (*Exp 3*), 74.5% of the subjects were male students, 25.2% females; subjects' age ranged from 19 to 47 years old (median=19 years). Finally, in the fourth experiment (*Exp 4*), 76.2% of the subjects were male students, and 23.8% were female students. The subjects' age ranged from 18 to 55 years old (median = 19 years). In terms of students' age, although the range of age varies in most experiments, our sample data showed a very small percentage of students in their 40's or 50's. In *Exp1*, there were 2 out of 48 (4%) students of age above 40 years old. In *Exp2*, there were 4 out of 214 (2%) students of age above 40 years old. In *Exp3* and *Exp4*, only 1 student was aged above 40 (out of 118 and 137 students respectively) representing less than 1% of the sample size. In *Exp5*, all students were below 32 years of age. In all experiments, more than 90% of students' ages were between 19 and 30s. With these data, we believe that those groups of older students did not impact our statistical analysis.

Table 2 Formal experiments characteristics

Experiment	*Exp 1*	*Exp 2*	*Exp 3*	*Exp 4*
Semester:	Summer 2009	Semester 1, 2009	Semester 2, 2009	Semester 1, 2010
Sample size:	48	214	118	137
Course:	CS101	CS101	CS101	CS101
Subjects:	First year undergraduate	First year undergraduate	First year undergraduate	First year undergraduate
Tutorial settings:	• Compulsory • 2 hours • Closed-lab	• Compulsory • 2 hours • Closed-lab	• Compulsory • 2 hours • Closed-lab	• Compulsory • 2 hours • Closed-lab
Personality factor (IV):	Conscientiousness	Conscientiousness	Neuroticism	Openness to Experience

In all experiments, subjects came from various ethnic backgrounds; the majority being the NZ/Pakeha, and Chinese. Other ethnic groups included South Korean, Asian, Indian, Middle Eastern, African, and Pacific Islanders. The samples used for our analysis were those who have consented to participate and completed the personality test and have taken either the midterm test or final exam. Table 2 summarizes the characteristics of each of the formal experiments.

In the first experiment (*Exp 1*), we hypothesized that there would be differences in performance between groups of paired students with similar and mixed Conscientiousness. In the second (*Exp 2*), we investigated whether different levels of Conscientiousness (low/medium/high) could have had an impact on paired students' academic performance. In *Exp 3*, we investigated whether differences in levels of Neuroticism (low/medium/high) when pairing had significant impact on students' academic performance. Finally, the fourth experiment (*Exp 4*) investigated the effects of Openness to experience on students' academic performance when pairing. In the following subsections we summarize the findings in terms of correlations between factors (both IV and DVs), and present an overall analysis based on the hypothesis testing of each experiment and the results from the post-experimental surveys. We included in the analysis the results from the power analysis which were not published before.

5.1 Findings on Correlation

Table 3 presents the aggregation of the bivariate Pearson correlation results between the three personality traits 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 two of the four 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' performance in the midterm test and final exam appeared to be mostly significantly and positively correlated with students' level of Openness to experience. In *Exp 3* and *Exp 4*, Conscientiousness showed a significant positive correlation with most academic performance criteria. Overall, paired students' academic performance was not associated with students' Neuroticism levels.

Table 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*	-0.03	-0.11	-0.08
	3	0.19**	0.19**	0.15
	4	0.17**	0.19**	0.18**
Neuroticism	1	-0.17	-0.04	-0.03
	2	0.02	-0.04	-0.04
	3*	0.05	-0.01	0.01
	4	0.04	-0.02	-0.00
Openness to	1	0.15	0.35**	0.29**
	2	0.21**	0.13	0.22**
	3	0.01	0.23**	0.15
	4*	0.15	0.18**	0.17**

N(*Exp 1*) = 48; N(*Exp 2*) = 214; N(*Exp 3*) = 118, N(*Exp 4*)=137
(*) Personality factor is controlled
(**) Significant at $\alpha < 0.05$

5.2 Hypothesis Testing

We used a single factor multivariate analysis (MANOVA) in *Exp 1* to analyze whether there was any significant difference in academic achievement between paired students of similar and mixed Conscientiousness levels. MANOVA 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 4 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 'N' column indicates the sample size. In *Exp1*, there were 22 pairs of similar Conscientiousness (i.e. homogeneous) and 21 pairs of mixed Conscientiousness.

Table 4 Mean and standard deviations (*Exp 1*)

DV	Personality Type	N	Mean	SD
Assignments	Similar Conscientiousness	22	13.07	2.08
	Mixed Conscientiousness	21	12.48	2.53
	Combined	43	12.78	2.30

Test Scores	Similar Conscientiousness	22	76.00	20.68
	Mixed Conscientiousness	21	83.83	16.21
	Combined	43	79.83	18.83
Final Exam	Similar Conscientiousness	22	73.11	18.68
	Mixed Conscientiousness	21	78.21	22.00
	Combined	43	75.60	20.29

The results of the Levene's Test indicate that the assumption of homogeneity of variances of each variable was not violated. Based on the Wilk's Lambda test generated by the MANOVA (see Table 5), results showed no significant differences (F =1.03, df=39, *p*=0.39) between the groups. 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**.

We conducted a post-hoc power analysis using the G*Power (3.1.2) to compute the power of the statistical test employed in our experiments (Faul et al., 2007). A statistical power represents the likelihood that a treatment effect will be observed whenever there is one. High power indicates greater ability to detect a difference between treatments if a true difference exists, when compared with a study with low statistical power (Dyba et al., 2006). In *Exp 1*, the power analysis results showed that our obtained power is considered to be low at the small effect size of 0.08. The power (1-β) indicates that the possibility of detecting a difference between the groups was only of 28% (See Table 5).

Table 5 Multivariate Test (*Exp 1*)

Test Approach	Value	F	Hypothesis df	Error df	*p* value	Effect Size	Observed power (1-β)
Wilks Lambda	0.93	1.03	3	39.0	0.39	0.08	0.28

In *Exp 2*, the null hypothesis was tested using the one-way analysis of variance (ANOVA) test to analyze whether there were any significant differences in academic performance between the three levels of Conscientiousness (low, medium, and high). ANOVA is chosen when there is only one independent variable with three or more levels and at least one continuous dependent variable (Pallant, 2007). It compares the variance between the groups of low, medium and high Conscientiousness 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 6 provides the mean values and standard deviation values for dependent variables for each group. We revised the data analysis

to address the issue of equal balance of data points more effectively. Hence mean values presented herein are slightly different from those given in Salleh et al. (2010a); however, the results and the conclusions drawn from those results remained unchanged. The overall *F* values for the three ANOVA are presented in Table 7.

These results show that there were no significant differences in academic performance between the three groups of Conscientiousness (at $\alpha = 0.05$). Thus, we could not find strong support to reject the null hypothesis (H1_O). The results indicate that **PP's effectiveness was not affected by differences in Conscientiousness levels among paired students**. The power analysis results for *Exp 2* showed that the effect size and the power of statistical test were considered to be very low. For instance, a power value 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). The analysis was carried out separately for each dependent variable using the F-test family of the one-way ANOVA.

Table 6 Mean and standard deviations (*Exp 2*)

DV	Personality Type	N	Mean	SD
Assignments (Range: 0 to 15)	Low Conscientiousness	70	11.49	3.99
	Medium Conscientiousness	74	11.60	4.25
	High Conscientiousness	70	12.11	3.63
	Combined	214	11.72	3.96
Test Scores (Range: 0 to 100)	Low Conscientiousness	70	83.56	17.13
	Medium Conscientiousness	74	81.17	20.56
	High Conscientiousness	70	82.05	20.43
	Combined	214	82.24	19.39
Final Exam (Range: 0 to 100)	Low Conscientiousness	68	75.06	19.34
	Medium Conscientiousness	72	73.64	20.36
	High Conscientiousness	69	73.32	21.53
	Combined	209	73.99	20.35

Table 7 ANOVA results (*Exp 2*)

DV	Sum of squares	df	Mean Squares	F	*p* value	Effect Size	Observed power (1-β)
Assignments	16.01	2	8.01	0.51	0.60	0.07	0.14
Test Scores	208.86	2	104.43	0.27	0.76	0.06	0.12
Final Exam	118.84	2	59.42	0.14	0.87	0.04	0.08

In *Exp 3*,the hypothesis was tested using the ANOVA (at $\alpha = 0.05$) in order to compare the effects of the three levels of Neuroticism on paired students' academic performance. Table 10 shows the values of the mean scores and standard deviations for each Neuroticism level. The Levene Test result showed that the assumption of homogeneity of data was not violated, thus population variances for each groups were not significantly different from each other. The

ANOVA results (see Table 9) 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).

Our results indicated that we could not find strong support to reject the null hypothesis (H2_O). Therefore, based on our data, we found that **paired students' academic performance was not significantly affected by differences in Neuroticism levels**. Our power analysis for *Exp 3* presented low statistical power relating to all the DVs. 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 11). With these effect sizes and the given sample size, the power generated was 0.47 and 0.54 respectively, which are 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).

Table 8 Mean and standard deviations (*Exp 3*)

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

Table 9 ANOVA Test (*Exp 3*)

DV	Sum of squares	df	Mean Squares	F	*p* value	Effect size	Observed power (1-β)
Assignments	133.39	2	66.69	2.45	0.09	0.20	0.47
Test Scores	2806.18	2	1403.09	2.93	0.06	0.22	0.54
Final Exam	2630.18	2	1315.09	1.80	0.17	0.18	0.36

In *Exp 4*, the null hypothesis was tested using ANOVA to analyze whether there was any significant difference in academic performance between the three levels of Openness to experience (low, medium, and high). Table 10 provides the mean and standard deviation values for academic performance for each group. 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 indicate that the

variances of scores were significantly different for each group of Openness (i.e. *p* 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). 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 (at $\alpha = 0.05$). 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 Mean and standard deviations (*Exp 4*)

DV	Personality Type	N	Mean	SD
Assignments (Range: 0 to 15)	Low Openness	48	11.02	3.58
	Medium Openness	47	10.24	3.41
	High Openness	42	12.06	2.23
	Combined	137	11.07	3.23
Test 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
	Combined	137	65.56	24.00
Final Exam (Range: 0 to 100)	Low Openness	48	66.83	26.46
	Medium Openness	46	60.17	28.51
	High Openness	42	79.30	19.35
	Combined	135	68.49	26.23

Table 11 Robust Test Equality of Means using Welch (*Exp 4*)

DV	F	df1	df2	*p* value	Effect size	Observed power (1-β)
Assignments	4.79	2	87.51	0.01	0.24	0.70
Test Scores	7.43	2	88.81	0.001	0.30	0.88
Final Exam	7.65	2	86.72	0.001	0.30	0.88

Post-hoc comparisons were performed to further examine for which groups the means differed. 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). The results from applying the Games-Howell test can be summarized as follows: i) **paired students of high Openness achieved better performance in assignments, midterm test, and final exam when compared with their counterparts**; and ii) **paired students of lower and medium Openness had comparable performance in assignments, midterm test, and final exam**. The power analysis

for *Exp 4* 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 (see Table 11).

5.3 Results from Quantitative surveys

Data on students' feedback about working with their partner was gathered using a questionnaire (see Appendix A). This was designed to measure 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 12 Summary of paired students feedback

	(% of Agree/Strongly Agree)			
Item/ Percentage (%)	***Exp 1***	***Exp 2***	***Exp 3***	***Exp 4***
Satisfaction level	88.5	90.2	85.7	87.2
Confidence level*	87.9	87.7	84.3	84.9
Productive Experience	90.4	95.0	90.0	89.7
Enjoyment	92.6	94.0	88.5	89.7
Increase Motivation level	86.0	87.0	84.3	83.9

(*) % indicates responses with High/Very High confidence

Table 12 presents a summary of the results. On average, **87.9% of students gained high satisfaction from the PP experience** and **86.2% responded that their confidence level in solving the programming exercises was high**. Likewise, **most students (91% on average) felt that PP was a productive experience, enjoyable (91%) and helped increase their motivation level (85% on average).**

6. Discussion

We discuss the key findings from our four experiments and the validity threats. The aggregation of the hypothesis testing results and the associated statistical power analysis of each experiment is presented in Table 13. The results from the first two experiments (*Exp 1* and *Exp 2*) indicate that there was a lack of evidence to differentiate performance of paired students based on their Conscientiousness levels. There was also a lack of evidence to support our alternative hypothesis on Neuroticism in *Exp 3*. Finally, we obtained evidence that supported our alternative hypothesis in *Exp 4* where the Openness to Experience trait significantly distinguished academic performance of paired students.

Each of the four formal experiments included a post-hoc analysis of statistical power to help interpret their results. The 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).

Table 13 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	214	Conscientiousness	No	ANOVA	0.07 (assign.) 0.06 (midterm) 0.04 (final)	0.14 0.12 0.08
3	118	Neuroticism	No	ANOVA	0.20 (assign.) 0.22 (midterm) 0.18 (final)	0.47 0.54 0.36
4	137	Openness to Experience	Yes	ANOVA	0.24 (assign.) 0.30 (midterm) 0.30 (final)	0.70 0.88 0.88

(*) Alpha (α) is set to 0.05 in all experiments

When investigating for the effects of Conscientiousness, the range of statistical power varied widely from 0.08 to 0.28. 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 4* 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 13).

In terms of the effect size, we observed that the effect sizes were remarkably low (i.e. between 0.04 to 0.08) when differentiating the performance of paired students based on students' Conscientiousness level in *Exp 1* and *Exp 2*. 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 to be the 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 improvement in academic achievement in assignments, midterm test, or exam. Note that the effect size estimated in our analyses was 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 has to be estimated from the sample data (Yuan & Maxwell, 2005).

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 *Exp 1* and *Exp 2* we could not find significant evidence to distinguish paired students' academic performance based on Conscientiousness levels. Similarly, *Exp 3* found paired students' performance was not significantly affected by the different levels of Neuroticism. However, the low power level exhibited suggests the patterns observed may or may not be likely to apply to other samples from the same population of interest.

Exp 4 results suggests that the level of paired students' Openness to experience could impact students' academic performance significantly more than their Conscientiousness and Neuroticism. This is in line with studies reported in the personality/educational psychology literature that observe the nature or

characteristics of open individuals as being bright, broad-minded and creative, which is as a consequence thought to eventually bring significant advantages for their academic success (Paunonen & Ashton, 2001; Farsides & Woodfield, 2003; Philips et al., 2003; Lounsbury et al., 2003). 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). Sound studies found 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). This means a team composed of higher aggregate levels of Openness to experience resulted into higher levels of knowledge sharing (Hsu et al., 2007). In the context of PP, while high Openness students obtain better performance when pairing, those who are low in Openness might benefit from pairing with someone who is medium or high in Openness. We believe that an interesting direction for future work relates to exploring whether PP helps encourage Openness, at least for students engaging in PP tasks. For instance, if a low Openness student were to be paired with a higher Openness student, would the pair work facilitate students to be more broad-minded or more diverse in thinking?

6.1 Threats to the validity of the findings

Threats are grouped based on four types of validity issues (Cook & Campbell, 1979): statistical conclusion validity, internal validity, constructs validity, and external validity.

6.1.1 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 2 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) claim 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 2 error. The low power observed in some of the experiments indicates that our data does not warrant the assumption 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 Conscientiousness, or Neuroticism.

Another threat relates to the violation of assumptions of statistical tests used in the experiments. In particular, for the *Exp 4*, 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 as 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 the appropriate *Games Howell* test 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).

6.1.2 Internal Validity

Internal validity threats are related to issues such as experimental procedures, treatments, or programming 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 as the sample could not be considered random. This "self-selected" sampling therefore is the main source of internal validity threat.

In terms of pair configuration employed the allocation of pairs was done randomly with respect to students' personality trait levels and automated by our PALLOC software. All participants were first or second year undergraduate students and their academic background appeared to be generally similar. Therefore, the potential for selection bias was minimized.

In our study, the senior tutor who lectured the tutorials for the introductory programming course (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 the low level of difficulty of the tutorial classes may have influenced our findings, so contributing for the non-significance of the findings. However, further work is needed in order to investigate the impact that task complexity has upon pairing effectiveness.

There is also a tendency for 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. Future replication studies should consider gender as a possible factor when investigating the effects of personality traits on PP.

The fact that the courses and tutorials employed in our experiments were taught or handled by several instructors and 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 these experiments.

Although students were aware of the experiment's objectives (i.e. from the *Participant Information Sheet*) and their own personality traits, they were not aware of the investigated hypothesis. Moreover, upon signing the consent form, students were informed that their participation was voluntary and that their decision whether to participate or not would 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. Surveys were also monitored by the tutor. These issues reduce the potential for social threats.

6.1.3 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 constructed a survey questionnaire intended to measure students' perception regarding their pairing experience in terms of satisfaction, confidence, and enjoyment levels. 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). The questionnaire was designed using a five-point scale to enable subjects to choose the answer that best represents their perceptions of the pairing experience. The surveys were distributed at the end of each tutorial and therefore the time spent on them was quite limited. Results showed that most 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. 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 may also be affected by third party variables such as learning strategies, cognitive ability, or self-motivation. One might presume that there was significant home studying between lab sessions. We believe that this did not really occur, and students relied mostly on what they were presented in the tutorials and in the classes. Since our study aimed to improve students' learning by practicing PP throughout a semester, measuring their academic performance was in our view appropriate for use 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., 2011).

The IPIP-NEO we 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 within the context of this research it was less likely for students to respond in socially desirable ways because they knew that their responses would not affect their academic record.

Classifying the personality scores into 30-40-30% could potentially contribute as a validity threat. According to Johnson (2008), the cut off points for low, medium, and high scores are arbitrary. There are no discernible differences between someone at the high end of "medium" and someone at the low end of "high.", hence suggesting that the boundaries are very fuzzy. In reality, personality traits are continuous, not trichotomous. Thus, this may be one of the reasons for not obtaining data that supports our hypotheses.

In the context of personality disorder, it was reported that there is a substantial relationship between the five factors and many of the personality disorder scales. Although the ability of the FFM to diagnose specific personality disorder categories is limited, the FFM dimensions can be used to differentiate personality disordered groups from general population (Saulsman & page, 2003). This is because the FFM was designed to diagnose normal range personality and that every personality disorder category is related to the FFM in some way (Saulsman & page, 2003). Based on this evidence, we believe that such criticisms about the FFM did not appear to significantly affect our study design and findings.

6.1.4 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 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, three of our four formal experiments (*Exp 1*, *Exp 2*, and *Exp 3*) 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 4* 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 that all experiments were conducted using subjects enrolled in an introductory programming course. The effects of personality on PP 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. This motivated us to replicate one of our experiments.

7. Replicated Study (*Exp 5*)

This Section describes a study (*Exp 5*) that replicated *Exp2*. Both took part during Semester 1, 2009. Exp 5 used as subjects 77 second-year students attending a Software Design course (COMPSCI 230). Similarly to Exp2, this study also investigated the effect of the Conscientiousness trait on PP's effectiveness. Conscientiousness was chosen because it has been shown in some previous studies to be the most relevant trait influencing academic success in tertiary education (Poropat, 2009; Pulford & Sohal, 2006). This replicated study also provided us a way to investigate whether the results would differ, when compared to *Exp 2*, when participants were second-year Computer Science students, and tasks were related to Software Design, databases and client-server computing, rather than solely introductory programming. COMPSCI 230 consisted of ten weeks of lectures and a weekly non-compulsory tutorial. The course comprised four major parts including software design using UML, object-orientation, database modelling, and JDBC programming. As part of their assignments, students were required to model software applications using UML and ER diagrams, design databases, and develop client applications in Java to manipulate data stored in a Relational database system.

Within the context of COMPSCI 230, closed-lab tutorials, lasting for one hour each, were prepared for students needing help in understanding the subject matter; hence attendance was not mandatory. Students intending to attend a tutorial were requested to inform the tutor prior to the session. This was to enable us to assign students into pairs. During tutorials paired students were given exercises, which

were not graded but were discussed at the end of each tutorial. Grades were determined based on the assignments, a midterm test, and final exam grades.

This replicated experiment investigated whether similar results to those obtained in *Exp 2* were produced when applying the same experimental setup to a different group of subjects and programming tasks. It also looked into the association between students' personality score with the level of satisfaction and confidence when working in pairs. The instruments and procedures of this replicated experiment were similar to the series of four experiments (see Section 4.4).

7.1 Data Analysis

The experimental subjects were second year undergraduate students, where 82.1% were male students and 17.9% were females. Subjects' age ranged from 19 to 32 years old (median = 21 years). Table 14 shows the results of applying the bivariate Pearson's correlation test to measure the association between personality traits and academic performance.

Table 14 Correlation between traits and academic performance (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 (2-tailed).

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

Similar to our findings in the second experiment (*Exp 2*), 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 *Exp 2*.

The hypothesis was tested using ANOVA in order to compare the three levels of Conscientiousness on paired students' academic performance. Table 15 shows the mean and standard deviation values for each group. Mean differences are very

similar for assignments and final exam grades, whereas means for the test grades varied between Conscientiousness levels. The results from the Levene test 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 exam).

Table 15 Mean and standard deviations (Replicated Experiment)

DV	Personality Type	N	Mean	SD
Assignments (Range: 0 to 20)	Low Conscientiousness	25	14.19	3.87
	Medium Conscientiousness	31	13.53	3.22
	High Conscientiousness	21	14.37	3.61
	Combined	77	13.97	3.52
Test Grades (Range: 0 to 15)	Low Conscientiousness	25	7.57	3.17
	Medium Conscientiousness	31	5.65	3.03
	High Conscientiousness	21	8.40	2.62
	Combined	77	7.02	3.16
Final Exam (Range: 0 to 65)	Low Conscientiousness	25	36.53	13.47
	Medium Conscientiousness	31	35.85	8.90
	High Conscientiousness	21	38.78	14.15
	Combined	77	36.87	11.95

The overall F values for the three ANOVA tests are presented in Table 16. These results showed that there was no significant difference between students' performance in assignments and final exam (at $\alpha = 0.05$). However, results showed a statistically significant difference between Conscientiousness levels based on the midterm grades. Post-hoc comparisons using Turkey HSD test indicated that the test's mean grade for high Conscientious group (M=8.4; SD=2.62) was significantly different from the test's mean grades for medium group (M=5.65; SD = 3.03; $p = 0.05$). Thus, based on our sample data, results indicated that **there were no significant differences between performance and Conscientiousness levels**, providing support for the null hypothesis (H1_0). The exception was on the midterm test, where significant differences were found, thus rejecting the null hypothesis. The results from the power analysis showed that the effect size and the power of the statistical test varied according to the observed DVs. 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.

Table 16 ANOVA Test (Replicated Experiment)

DV	Sum of squares	df	Mean Squares	F	*p* value	Effect size	Observed power (1-β)
Assignments	10.74	2	5.37	0.43	0.66	0.11	0.12
Test Scores	105.65	2	52.83	5.98	0.01	0.32	0.71
Final Exam	111.71	2	55.86	0.38	0.68	0.10	0.11

Our data analysis showed that on average 79% of students attending the tutorials were satisfied working with their partner and 75% responded that their level of confidence solving the tasks with their partner was high.

7.2 Comparison of findings between *Exp 2* and *Exp 5*

In the replicated experiment, the effect of different Conscientiousness levels on academic performance was investigated using a more advanced CS course attended by more mature students, compared with *Exp 2*. However, similar to the findings from *Exp 2*, we did not observe in *Exp 5* any significant correlation between the personality trait Conscientiousness and academic performance. Our hypothesis testing indicates that results differed significantly only for the midterm test grades, where paired students of high Conscientiousness levels achieved higher grades 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 the midterm test. This suggests that the effect of Conscientiousness on paired students' academic performance may be trivial regardless of the nature of the courses. The overall low statistical power observed in this experiment is similar to the power generated from *Exp 2* (See Table 17).

Table 17 Comparison of findings between *Exp 5* and *Exp 2*

Exp.	N	Personality Factor	Supported Hypothesis? (Yes/No)	Statistical Test (*)	Effect Size	Statistical Power
Exp 2	214	Conscientiousness	No	ANOVA	0.07 (assign.) 0.06 (midterm) 0.04 (final)	0.14 0.12 0.08
Exp5	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

(*) Alpha (α) is set to 0.05 in both experiments

In the replicated experiment the statistical power for the midterm ANOVA test (0.71) was found to be reasonably high due to the medium effect size (0.32) observed for this particular test. 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 influenced the results and more conscientious 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 (see Table 14). 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 for experiments *Exp 1* and *Exp 2*, and they also corroborate results reported in the educational-psychology literature (Farsides & Woodfield, 2003; Chamorro-Premuzic & Furnham, 2008). We were not able to replicate the experiments that investigated the Neuroticism and Openness to experience traits due to time constraints, but these replications are part of our future research plans.

8. Implications of our Findings

8.1. Implications for Researchers

Based on the aggregation of the results from our experiments, several implications for researchers can be drawn. We believe there are variables that could potentially mediate the personality-performance relationship in PP. Nevertheless they were not investigated within the context of our work. It is however our view 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). We suggest future studies should look into possible effects of mediator variables to gain insight into the mechanism underlying the personality-performance relationship in PP.

In all of the formal experiments we conducted, the effects of personality traits were investigated from the perspective of a broader-level or higher level personality trait. Each personality trait 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). Thus, an extension of this research could be to investigate 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). We did not use

the more detailed personality facets in the experiments due to a limitation in sample size.

Bowers et al. (2000) observe that team performance could be affected by task 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). All of our four 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. Results from the replicated study showed mixed findings with regard to the impact of Conscientiousness of paired students attending a more advanced computing course. We believe future work should investigate effects of personality on PP's effectiveness for more difficult tasks.

Due to a limitation in sample size, each experiment we conducted investigated only a single FFM personality trait. Students' academic performance may also be affected by another personality factor in the FFM or other non-personality variables such as intelligence, skill level, or gender. Nguyen, et al. (2005) reported 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 a larger sample size should be employed in future studies to determine if there is any interaction effect between personality and other factors that may potentially affect paired students' academic performance.

Further research might also explore the issue of personality in PP using a qualitative approach such as a case study, ethnographic study, grounded theory or content analysis. Qualitative investigations may facilitate in further deepening our understanding of the research results by collecting various forms of data and portraying the issue in its multifaceted form.

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. 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 been suggested that studies

should perform a priori power analysis to obtain estimates of sample size expected to achieve a high statistical power (Lan & Lian, 2010). Nevertheless, we were constrained by the class sizes available, which could not be increased.

The two other personality factors also merit investigation. Agreeableness is the personality trait that 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. Extraversion is a trait that 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. Extraversion was also positively correlated with software quality and software teams with a higher aggregate on Agreeableness achieved the highest job satisfaction (Acuna et al., 2009). Research into these factors may result in a better understanding of their influence on PP's effectiveness in higher education.

8.2. Implications for Educators

Our findings imply that pairing students according to either conscientiousness or neuroticism levels do not appear to be significantly important in ensuring successful academic performance of paired students. Therefore, PP group formation and monitoring by CS/SE educators may be able to ignore the pairing formation based on personality trait conscientiousness and neuroticism for such an introductory programming course. Our empirical evidence showed mixed findings with regard to the impact of conscientiousness levels of academic performance of paired students attending a more advanced Computing course. Thus, it would be necessary to conduct a further study that involves more complex tasks to determine whether task difficulty level plays a significant role in differentiating paired students' academic performance based upon their personality traits.

Our results showed a greater performance of high Openness students than those of lower Openness. Farsides and Woodfield (2003) note that "*being Open to experience provides academic benefits beyond those provided by being clever and*

being motivated to turn up to classes" (p. 1239). Thus, we believe that this trait may probably be the most 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.

One practical implication is that PP does not appear to give harmful effects to either students' satisfaction or confidence level in an introductory programming course. Our findings indicate students' motivation, enjoyment, satisfaction, and confidence level were very encouraging regardless of their differences in personality trait, consistent with other findings in the PP literature (DeClue, 2003; Hanks, 2006). This supports educators in continuing to employ PP as a pedagogical tool in an introductory learning to program course.

9. Conclusions and Future Research

We found evidence suggesting that differing levels of the personality trait Conscientiousness would not affect the academic success of paired students in CS introductory programming courses. Our results are counterintuitive to many of the findings reported previously in the educational-psychology literature that Conscientiousness is significantly positively correlated to students' academic performance (Poropat, 2009; Busato et al., 2000). It is important to note that given a lack sufficient statistical power to detect effects of interest we not generalize our results to the wider CS/SE population. 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 in a more advanced course. Thus, it would be necessary to conduct further studies involving more complex tasks to confirm whether task difficulty level plays a significant role in differentiating paired students' academic performance based upon their personality traits.

Similarly, differences in Neuroticism levels in pairs (low/medium/high) were found not to affect significantly paired students' academic performance. Once again the low statistical power obtained prevents us from concluding the effects of this personality trait on students' academic performance. This lack of statistical power means that the possible effects of Conscientiousness and Neuroticism cannot be ruled out. Under such 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 and Neuroticism are indeed

absent in the target population. A positive finding might be obtained in a future study if an adequate sample size or more sensitive research design are employed.

Conversely, we found Openness to experience had a substantial impact towards paired students' performance, where paired students consisting of high Openness achieved significantly better academic performance compared with their counterparts. 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.This indicates its practical significance or importance for distinguishing students' academic performance. 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.

Our findings also indicate that despite the variation in students' personality profiles 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. These findings shed some light on our understanding of the influence of personality traits in PP from the perspective of the FFM. We recommend future replication studies to investigate the effects of personality traits Conscientiousness, Neuroticism, and Openness on paired students' academic performance in order to confirm or refute the findings we obtained from this research. We also believe that future work should investigate whether the personality traits of pairs actually do impact upon the performance of design or testing tasks, or on tasks of higher difficulty level than those from an introductory CS programming course.

Aggregating the results of our experiments were applicable within the context of undergraduate students' learning in an introductory programming course. However, when replicating our second experiment, we obtained one significant correlation between the level of Conscientiousness and the mid-term grades. This suggests further research is needed to investigate whether findings converge or diverge when employing senior level students, and tasks that are not only programming-related. In addition, performing a qualitative study in the future may be practical in order to better understand the results obtained in the present study.

Acknowledgements

This research is supported by the Ministry of Higher Education Malaysia. The authors would like to thank all tutors and demonstrators of CS101 and CS230 at the University of Auckland (2008-2010) for the help given to run the experiments. Thanks also to all students who have participated in the experiments.

References

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.

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.

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.

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.

Beck, K. (1999). *Extreme Programming Explained: Embrace Change (*2nd Ed.). Boston, US: Addison-Wesley.

Bell, S.T. (2007). Deep-level composition variables as predictors of team performance: A Meta-Analysis. *Journal of Applied Psychology, 92*(3), 595-615.

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.

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.

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.

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.

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.

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

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.

Conard, M.A. (2006). Aptitude is not enough: How personality and behavior predict academic performance. *Journal of Research in Personality, 40*, 339 - 346.

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. (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.

Cunha, A.D.D., &Greathead, D. (2007). Does personality matter? An analysis of code-review ability.*Communications of the ACM, 50*(5), 109-112.

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.

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., 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., 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.

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., 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.

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.

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. (1996). 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.*Proc. of the International Symposium on Empirical Software Engineering (ISESE'03)*, 132-142.

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.

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.

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.

Hicks, L. (1984). Conceptual and empirical analysis of some assumptions of an explicitly technological theory. *Journal of Personality and Social Psychology*, 46, 1118 – 1131.

Ho, C.-w. (2004). Examining impact of pair programming on female students (No. TR-2004-20). Raleigh, NC: North Carolina State University.

Hsu,, B.-F., Wu, W.-L., and Yeh, R.-S. (2007). Personality composition, affective tie, and knowledge sharing: A team level analysis. *Proceedings of the Portland International Center for Management of Engineering and Technology (PICMET 2007)*, 2583 – 2592.

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. (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/

Juristo, N., & Moreno, A. M. (2001). *Basics of Software Engineering Experimentation*. Boston: Kluwer Academic Publishers.

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.

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.

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.

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.

Levine, J.M., & Moreland, R.L. (1990). Progress in small group research. *Annual Review of Psychology*. 41, 585 – 634.

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.

Matzler, K., Renzl, B., Muller, J., Herting, S., & Mooradian, T.A. (2008). Personality traits and knowledge sharing, *Journal of Economic Psychology,* vol 29, 301-313.

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.

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.

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.

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, J.L., & Well, A.D. (2003). *Research Design and Statistical Analysis* (2 ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

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).

O'Connor, M.C., &Paunonen, S.V. (2007). Big five personality predictors of post-secondary academic performance.*Personality and Individual Differences, 43*, 971-990.

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.

Pfleeger, S.L. (1995). Experimental design and analysis in software engineering.*Annals of Software Engineering, 1*(1), 219-253.

Philips, P., Abraham, C., & Bond, R. (2003). Personality, cognition, and university students' examination performance. *European Journal of Personality*, 17, 435 – 448.

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.

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.
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. (2011). Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review. *IEEE Transactions on Software Eng.,* 37(4), 509-525
Salleh, N., Mendes, E., and Grundy, J. (2011a). The Effects of Openness to Experience on Pair Programming in a Higher Education Context*,* in *Proceedings of the 24th IEEE-CS Conference on Software Engineering Education and Training (CSEET2011*), Honolulu, Hawaii, IEEE Computer Society.
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)*.
Saulsman, L.M and Page, A.C. (2003). Can trait measures diagnose personality disorders? *Current Opinion in Psychiatry*, Volume 16(1), pp 83-88.
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(2), 187-226.
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).
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.
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.
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.
Yuan, K., & Maxwell, S. (2005). On the Post Hoc Power in Testing Mean Differences. *Journal of Educational and Behavioral Statistics, 30*(2), 141-167.

Appendix A. PP 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).

Q8. How do you rate your level of confidence solving the exercises with your partner? (Circle only one)

Q9. Comments:

Thank you for your time!