An Empirical Study of the Effects of Personality on Software Testing

Tanjila Kanij  
Swinburne University of Technology  
Hawthorn Australia  
tkanij@swin.edu.au

Robert Merkel  
Monash University  
Clayton Australia  
robert.merkel@monash.edu

John Grundy  
Swinburne University of Technology  
Hawthorn Australia  
jgrundy@swin.edu.au

Abstract

The effectiveness of testing is a major determinant of software quality. It is believed that individual testers vary in their effectiveness, but so far the factors contributing to this variation have not been well studied. In this study, we examined whether personality traits, as described by the five-factor model, affect performance on a software testing task. ICT students were given a small software testing task at which their effectiveness was assessed using several different criteria, including bug location rate, weighted fault density, and bug report quality. Their personality was assessed using the NEO PI-3 personality questionnaire. We then compared testing performance according to individual and aggregate measures against different five-factor personality traits. Several weak correlations between two of these personality traits, extraversion and conscientiousness, and testing effectiveness were found.

1 Introduction

In our recent survey of professional testers [34], we found that the recipients strongly believed that there are substantial individual differences in tester effectiveness. Given the effort devoted to software testing, and its importance to the quality of software systems, it is natural to investigate the nature of these individual differences. A better understanding of these factors could assist in the identification of individuals with high potential to be good testers, and inform training and development of software testers.

Our respondents indicated that “domain knowledge”, “experience”, “general intelligence” and “different personal characteristics” contributed significantly to this variance. Some of these characteristics are relatively straightforward, and common across most skilled professions. The “general intelligence” commonly assessed using IQ tests, is positively correlated with job performance in a variety of professions [37], though the basis of this correlation is controversial [33]. Similarly, it is unsurprising that experience and knowledge in the problem domain improve testing performance. Beer and Ramler [6] found that experienced testers had increased domain knowledge that is helpful to understand incomplete and ambiguous specifications. “Different personal characteristics”, is however, a less straightforward notion. We believe that these “characteristics” that our respondents referred to, include individual personality traits. Unlike “intelligence” or “domain knowledge”, it is not at all clear how particular personality traits might relate to being an effective software tester.

The empirical study reported in this paper was been designed to investigate the effect of personality traits, assessed according to the five factor model of personality [32], on effectiveness of software testing. The study collected the personality profiles of ICT students and measured their effectiveness in a software testing task. We then determined whether there were any relationships between personality traits and effectiveness. A custom software testing task was designed to assess the effectiveness of students in software testing in our study. Personality profiles were collected using the NEO personality inventory test (NEO PI-3). If specific personality traits have influence on
testing effectiveness, these traits could be used to aid recruitment and selection of software testers. This could also help young IT graduates to select the most appropriate IT career for them.

The rest of the paper is organized as follows: Section 2 discusses background and related work of this research study, Section 3 describes the details of this empirical study, Section 4 presents the findings of this study, Section 5 indicates what can threaten the results, Section 6 discusses the details of the findings and finally Section 7 concludes the paper with possible future work.

2 Background and related work

2.1 Personality

While “personality” is a very widely-used concept, a universally-accepted formal definition of the concept remains elusive. Generally, personality is used to refer to enduring non-intellectual personal characteristics that differentiate the individual from others. One standard method of describing a personality is in terms of traits, which constitute “...very generalized behaviour patterns in response to emotional tendencies” [32].

There have been many attempts to find a manageable small set of words that represent major personality traits. A result of such attempts is the “big five” factor model [32]. The big five factors are extraversion, agreeableness, conscientiousness, neuroticism and openness to experience. The “big five” factor model has been widely used in personality psychology.

A particularly popular tool for personality assessment research is the Myers Briggs Type Indicator (MBTI), developed based on the theories of Carl Jung [32]. The basic inventory encompasses two kinds of functions and four preference dichotomies - Sensing(S) - Intuition(N), Thinking (T) - Feeling (F), Extraversion (E) - Introversion (I) and Judging (J) - Perceiving (P). The permutation of the four preference dichotomies results in sixteen personality types forming MBTI inventory.

2.2 Research on personality factors related to software testing and debugging

While most software testing research to date has focused on the technical side of the discipline, there have been a few studies that have examined the connection between ability in aspects of testing and various factors relating to personality. Da Cunha and Greathead [18] found that logical and ingenious people, as measured by the MBTI, were good at code review tasks. Almodaimeegh et al. [3] assessed the “locus of control” – that is, the extent to which one believes that an event is the result of external factors than his/her own effort – of programmers. They also investigated the influence of social learning style on debugging skill. They found no significant relation between locus of control and debugging skill, however, individual social learning style and experience were influential on debugging performance. Similarly Shoaib et al. [40], studied the effect of personality traits of students assessed with MBTI, on the effectiveness of exploratory testing and concluded that extravert personality traits are positively correlated with effective exploratory testing. However, whether extraverts will, in general, be good testers still remains an open question.

2.3 Research on other human factors related to software testing

The influence of attributes other than personality, such as experience and work organization, has also been investigated in many works. Beer and Ramler [6] and Juha et al. [26] conducted industrial case studies in which they found that more experienced testers used domain knowledge to interpret incomplete specifications, and chose suitable techniques for testing. Organisational issues in testing have also attracted relatively little attention. The limited studies to date [39, 38, 30], however, indicate that attitude and organisational factors are significant, as testing is considered a boring task and many difficulties relate to organizational factors rather than technical ones. Working within an
organization, and the ability to cope with a sometimes very monotonous task, are likely to relate to personality traits.

2.4 Research on personality factors related to programming

There has been a considerable body of research investigating personality types of programmers. Several early studies investigated whether specific MBTI personality traits were over-represented in the programming community, [10, 29]. More recent research examined the impact of these personality types on performance. Cegielski and Hall [14] found that personality type was a predictor of performance in OO programming. However, Darcy and Ma [19] assessed the influence of personality, as measured using the five-factor model, on the performance of student programmers and found no effects.

Other software development tasks have also been examined in this manner. Arockiam et al. [5] investigated the connection between gender, personality traits, and object oriented code comprehension skill. They found that male student participants with cooperativeness and high emotional ability did perform well. Greathed [24] found that introverted people are good at code comprehension. Ahmed et al. [2] found that judging and thinking type students performed better in a software design course.

2.5 Research on personality factors related to software engineering

There has been some research analysing the personality traits of software engineers in general. Capretz [11] administered the MBTI personality assessment test to 100 software engineers ranging from student to professional level. They found that sensing and thinking (ST), thinking and judging (TJ) and intuitive and thinking (NT) types were over-represented in the sample. Sodiya et al. [41] proposed a general test that assesses the five personality factors along with cognitive style with a standard questionnaire and suggests the software engineering role that best suits the assessed personality type. Feldt et al. [21] assessed the personality of 47 professional software engineers, using the IPIP inventory, and found that there are different clusters of personalities among them and each cluster have significant correlation with attitude. Clark et al. [15] assessed the personality of a number of IT students and professionals with Adjective Check List (ACL) and found that both exceptional students and professionals are conscientious, however, exceptional students are introvert while exceptional IT developers are extravert. The difference in personality type between exceptional students and developers raises the possibility that exceptional students might not become exceptional developers.

2.6 Expert views

While there is currently insufficient empirical evidence to make definitive statements on the relationship between personality and testing effectiveness, a number of experts in the area have expressed their opinion on the topic. According to Armour [4], good testers have a “nose for testing” - an intuition that helps them to determine what and how to test. Pettichord [35] listed a number of distinct characteristics of programmers and testers. According to him, testers should tolerate tedium, be skeptical and be comfortable with conflicts. Pol [31] suggests that testers should be creative, accurate and strict in their methodical approach. Black [7] thinks that testers are professional pessimists who possess the curiosity of looking for faults. Weyuker et al. [42] have outlined general and technical skills required for a tester. Capretz and Ahmed [12, 13] collected the job responsibilities mentioned in different job advertisements for different software engineering roles which also includes software testing and connected those responsibilities with required “soft” skills. Based on their experience and perception, the authors suggested that testers should have
attention to detail and should possess good organizational skills, which are common in sensing and judging type people as categorized by MBTI.

The conceptions of these testing professionals suggest that effective software testers have some specific characteristics. However, the empirical evidence supporting links to directly measurable personality characteristics is inconclusive, and in some cases, contradictory.

3 Study design

The goal of this study was to empirically determine relationships between specific personality traits and performance in software testing by software engineering and computer science students, if they exist. This section describes the design of a quasi experiment, in which student participants completed two tasks, an assessment of their personality and an assessment of their effectiveness in software testing.

3.1 Assessment of personality

MBTI has been commonly used in research relating personality and IT practitioners. However, neither the MBTI, nor Jungian personality theories, on which the axes of the MBTI are based, are widely used in modern personality psychology research. A major reason is because the MBTI does not come with norms based on continuous scores, restricting the scope of statistical analysis [9].

We have, instead, chosen to use the NEO personality inventory test (NEO PI-3), proposed by Costa and McCrae [17], and based on the five factor model of personality. In comparison to MBTI, NEO personality inventory tests are widely accepted by personality psychology community, and provide continuous normed scores on the standard five-factor model of personality. The five factors in NEO PI-3 are:

- **Extraversion (E)**: Extraversion is related to sociability, assertiveness, talkativeness and activity.
- **Agreeableness (A)**: Agreeableness encompasses the expressive quality of admirable human aspects of personality.
- **Conscientiousness (C)**: Conscientiousness is referred to as “Will to achieve” by Digman and Takemoto-Chock [20]. Conscientious people are purposeful, strong-willed and determined.
- **Neuroticism (N)**: This factor covers forms of excessive emotionality. Facets of this include anxiety, angry hostility, depression, self-consciousness, impulsiveness and vulnerability.
- **Openness to Experience (O)**: Openness to Experience is associated with intelligence and intellectual interests.

3.2 Assessment of effectiveness in software testing

3.2.1 Software Testing Task

Unfortunately, there is no standardized method or test for assessment of effectiveness in software testing. In order for us to assess the effectiveness in software testing in this study we have asked students to test a carefully crafted faulty Java program that calculates all possible loop free network topologies given the location of different towers defined with latitude and longitude and find the shortest topology among all. This Java program was originally developed as an assignment as part of an introductory “Programming in Java” unit in Swinburne University of Technology. A number of de-identified student assignments were collected and tested by the researchers. There were 18 different types of bugs encountered in those programs. We chose this program because both the program and the specifications were of manageable complexity and could thus be understood by participants in a reasonable time.
We selected the least buggy student assignment submission as the basis for the testing task in this study. The program used for our testing task had 1017 lines of code and the maximum cyclomatic complexity of the program was 7. Thirteen bugs found in other student assignments, along with 7 more bugs (which were not found in the student programs) were injected in the assignment program to be used for the study. Thus the testing task program had a total of 20 known bugs in it. The main reason behind injecting detected bugs is that the bugs represent real world bugs as they were created by the student programmers writing the program. Additional bugs were added since the student submissions contained insufficient “obvious” bugs for our purposes. The severity of the injected bugs was decided on the classification taken from [25]. This scheme was chosen because it is widely used and the definition of the categories is clear and straightforward.

To ensure that the injected bugs represent a reasonable approximation of real world bugs, we compared the type of bugs with two standard bug classification scheme, Knuth’s errors and Eisenstadt’s bug war stories [36]. Table 1 presents the comparison along with the description of the bugs.

<table>
<thead>
<tr>
<th>No.</th>
<th>Bug No.</th>
<th>Correct program behaviour</th>
<th>Incorrect program behaviour (bug)</th>
<th>Fault description</th>
<th>Fault Source</th>
<th>Severity</th>
<th>Knuth’s errors</th>
<th>Eisenstadt’s bug war stories</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The output is printed in the console and in the file</td>
<td>The output is printed only in the console</td>
<td>Missing function-ality</td>
<td>Students</td>
<td>Low</td>
<td>forgotten function</td>
<td>unsolved</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Pair of cities are printed with their names</td>
<td>Pair of cities are printed with names</td>
<td>Wrong implementation</td>
<td>Students</td>
<td>Medium</td>
<td>algorithm away</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Correct right number of pairs possible among the given cities</td>
<td>Print wrong number of pairs or cities</td>
<td>Wrong implementation</td>
<td>Researchers</td>
<td>High</td>
<td>found or botched</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Correct right distance between the pair of cities</td>
<td>Print wrong distance between pair of cities</td>
<td>Wrong function call</td>
<td>Researchers</td>
<td>High</td>
<td>mismatch between modules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Print pairs with right (number) of cities</td>
<td>Print pairs with wrong (number) of cities</td>
<td>Wrong function call</td>
<td>Researchers</td>
<td>High</td>
<td>found or botched</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Print right topology distance</td>
<td>Print wrong topology distance</td>
<td>Wrong index</td>
<td>Students</td>
<td>High</td>
<td>found or botched</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Selection of right shortest distance</td>
<td>Selection of wrong shortest distance</td>
<td>Wrong parameter</td>
<td>Students</td>
<td>High</td>
<td>found or botched</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Not printing duplicate topologies</td>
<td>Printing duplicate topologies</td>
<td>Wrong display</td>
<td>Researchers</td>
<td>Medium</td>
<td>found or botched</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Right display of distance unit</td>
<td>Wrong display of distance unit</td>
<td>Wrong value</td>
<td>Researchers</td>
<td>Medium</td>
<td>invavl type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Should calculate distance for all values</td>
<td>Cannot calculate topologies for special value (gives exception with 2 cities)</td>
<td>Students</td>
<td>Critical</td>
<td>Forgotten function</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Calculate right distance for cities with &lt;0 degree</td>
<td>Calculate wrong distance for cities with &lt;0 degree</td>
<td>Missing brackets</td>
<td>Students</td>
<td>High</td>
<td>invalid type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Should work with any number of cities</td>
<td>For more than 6 cities gives memory limit exceed exception</td>
<td>Memory</td>
<td>Students</td>
<td>High</td>
<td>invalid type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Right value of conditional operator in line 286 wirelessNet.java</td>
<td>Use of wrong conditional operator in line 286 wirelessNet.java</td>
<td>Wrong implementation</td>
<td>Researchers</td>
<td>High</td>
<td>invalid type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Use of right conditional operator in line 309 wirelessNet.java</td>
<td>Use of wrong conditional operator in line 309 wirelessNet.java</td>
<td>Wrong implementation</td>
<td>Researchers</td>
<td>High</td>
<td>invalid type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Calculate right distance for cities with 0 degree</td>
<td>Calculates wrong distance for cities with 0 degree</td>
<td>Missing equal sign</td>
<td>Students</td>
<td>High</td>
<td>invalid type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Should print error message when number of parameters mismatch</td>
<td>No output when number of parameters mismatch</td>
<td>Number of parameter mismatch</td>
<td>Students</td>
<td>Critical</td>
<td>surprising scenario</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Should print error message when complex parameter</td>
<td>No output when complex parameter</td>
<td>Parameter type mismatch</td>
<td>Students</td>
<td>Critical</td>
<td>surprising scenario</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Should print error message when parameter types mismatch</td>
<td>No output when parameter types mismatch</td>
<td>Parameter type mismatch</td>
<td>Students</td>
<td>Critical</td>
<td>surprising scenario</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Right method name 'get topology'</td>
<td>Wrong method name 'get topology'</td>
<td>Wrong implementation</td>
<td>Students</td>
<td>Low</td>
<td>invalid type</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.2.2 Assessment metrics

There are a number of proposed metrics in the literature for assessing testing effectiveness [22, 28]. Three metrics were used for our purpose: Bug location rate, Weighted fault density, Bug report quality. Given that there are no universally-accepted criteria for tester effectiveness, we judged that these three metrics were plausible and practical for use in our experiment.

- **Bug Location Rate**: This is measured as the number of bugs found in comparison to the time taken for testing [25].

  \[
  \text{Bug Location Rate} = \frac{\text{Number of bugs found}}{\text{Time taken for testing (in minutes)}}
  \]
Bug location rate could range from 0 to infinite.

- **Weighted Fault Density:** The weighted fault density can be measured as:

  \[
  \text{Weighted Fault Density} = \frac{(W_1 \times S_1) + (W_2 \times S_2) + (W_3 \times S_3) + (W_4 \times S_4)}{N}
  \]

  where:
  
  - \( W \) denotes assigned weights \( W_1 > W_2 > W_3 > W_4 \).
  - \( S \) denotes severity of faults assigned according to [25].
  - \( N \) denotes number of total bugs found.

  There were 4 critical, 11 high, 3 medium and 2 low bugs in the software. For this experiment we used the following weights: \( W_1 = 0.4, W_2 = 0.3, W_3 = 0.2 \) and \( W_4 = 0.1 \). The total weighted fault density could range from 0 to 0.4.

- **Bug report quality:** This is assessed using the IEEE standard of Test Documentation [1]. According to the standard, there are 8 separate documents that should be given as part of a bug report. As our test sessions were small, as well as the provided software to test was also a small and simple one, we did not expect the participants to provide separate document(s) on each of the bugs found. We examined whether any information on each of the 8 standard deliverable fields were present for each of the bugs reported in the bug report prepared by a student and assigned a score from 0 to 1 for each of the field. Thus the total score of bug report quality could range from 0 to 8.

The total number of bugs found was calculated as the number of reported valid bugs (Total number of reported bug - total number of reported invalid/false bug).

### 3.2.3 Assessment of overall effectiveness

Unfortunately in the testing literature there is no indication of how to assess overall effectiveness using the above defined metrics. For assessment of overall effectiveness using these metrics we followed two possible approaches:

- **Total score:** In this approach we summed the score of bug location rate, weighted fault density and bug report quality of a participant to derive overall score. The score is denoted as \( O_{\text{sum}} \).

  \[
  O_{\text{sum}} = \text{bug location rate} + \text{weighted fault density} + \text{bug report quality}
  \]

- **Weighted total score:** In this approach we assigned different weights to the assessment metrics according to their importance on total score to get an overall score. According to Kaner [27] the number of bugs found by a tester may be affected by a variety of factors. He emphasized qualitative assessment of bug report quality than accounting for bug counts [28]. In our previous study [34], we found that ‘Quality of bug report” is more important than “Severity of bugs found”. Therefore, we assign more importance to bug report quality. The other two proposed metrics are based on bug count and we can not distinguish the varying importance of those. We assign the following weights to the three proposed metrics (bug location rate = 0.3, weighted fault density = 0.3 and bug report quality = 0.4) and sum the weighted metrics to get weighted total score, denoted as \( O_{w\text{sum}} \).

  \[
  O_{w\text{sum}} = (0.3 \times \text{bug location rate}) + (0.3 \times \text{weighted fault density}) + (0.4 \times \text{bug report quality})
  \]

### 3.2.4 Validation of the instrument

We conducted a pilot study where our testing effectiveness measurement instrument was given to 9 volunteer student participants. After the pilot study, minor changes were made in the description of the specification document provided with the testing task.
3.3 Participants

The participants in this study were Computer Science and Software Engineering students of the Faculty of Information and Communication Technologies at Swinburne University of Technology and the Faculty of Information Technology at Monash University. Both universities are located in Melbourne, Australia. The researchers attended lectures for ICT undergraduate and postgraduate coursework units at both universities to invite students to participate.

3.4 Experimental procedure

After each of the participants signed a consent statement, they were assigned a unique number and were provided with a small demographic questionnaire asking the participants’ closed-form “yes/no” questions on whether they had any professional experience in software testing, any experience using testing tools, any professional experience in other IT areas, and whether they had completed any university units on software testing. Once the demographic questions were completed the participants were presented with the self scoring NEO PI-3 form. After the completion of the personality assessment form the participants were given the specification document of the testing task program. They were asked to verify if the program meets the specification and report any deviations as bugs. The debugging environment used was Eclipse (version 3.5.0) which is a standard IDE used by both universities. No testing technique was specified to the participants. At the end of the test each participant received $25 for participation. The T scores and the answer form was retained by the researchers. The participants were provided with their personality profile form named “Your NEO Summary”. The data was collected over the period September 2010 to October 2011.

4 Results

4.1 Participant demographics

A total of 48 students from 18 to 35 years of age, participated in the research study. The majority (69%) of participants were male. 22.3% participants had professional experience in software testing. 31.3% had studied a specialized unit on software testing. 27.1% had experience of using testing tools and 37.5% had experience in other areas of IT.

4.2 Population distribution

The scales of NEO personality inventory tests measure factors that follow an approximately normal distribution in the broader population. We have applied Shapiro-Wilk Test [8] and the results indicate that our population distributions do not differ significantly from normality.

4.3 Scores on personality traits and effectiveness in testing

Table 2 shows the minimum, maximum, average and standard deviation of the T-scores on five personality factors and the scores on the effectiveness measurement metrics obtained by our participants. The T scores are calculated based on the norms provided with the NEO PI-3 personality assessment test.

4.4 Correlation between personality traits and effectiveness in testing

Table 3 shows Pearson correlations between the five personality factors measured, and our testing effectiveness metrics. We found, at significance level $p < 0.05$, there is a negative association between extraversion and $O_{sum}$, $O_{osum}$ and bug report quality. We also see a positive, though weak association between conscientiousness and bug location rate and a weak negative association between conscientiousness and weighted fault density.
### Table 2. Distribution of scores (N = 48)

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Average</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroticism</td>
<td>32</td>
<td>74</td>
<td>54.94</td>
<td>10.38</td>
</tr>
<tr>
<td>Extraversion</td>
<td>20</td>
<td>70</td>
<td>50.42</td>
<td>9.32</td>
</tr>
<tr>
<td>Openness to experience</td>
<td>35</td>
<td>80</td>
<td>54.36</td>
<td>10.02</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>26</td>
<td>78</td>
<td>48.27</td>
<td>9.62</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>27</td>
<td>66</td>
<td>47.25</td>
<td>8.62</td>
</tr>
<tr>
<td>Osum</td>
<td>0.63</td>
<td></td>
<td>0.74</td>
<td>0.34</td>
</tr>
<tr>
<td>Owsum</td>
<td>0.31</td>
<td></td>
<td>0.73</td>
<td>0.34</td>
</tr>
<tr>
<td>Bug Location Rate</td>
<td>0.02</td>
<td></td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td>Weighted Fault Density</td>
<td>0.1</td>
<td></td>
<td>0.23</td>
<td>0.07</td>
</tr>
<tr>
<td>Bug Report Quality</td>
<td>0.5</td>
<td>3.5</td>
<td>1.56</td>
<td>0.83</td>
</tr>
</tbody>
</table>

### Table 3. Correlations (N = 48)

<table>
<thead>
<tr>
<th></th>
<th>Neuroticism</th>
<th>Extraversion</th>
<th>Openness to experience</th>
<th>Agreeableness</th>
<th>Conscientiousness</th>
<th>Osum</th>
<th>Owsum</th>
<th>Bug Location Rate</th>
<th>Weighted Fault Density</th>
<th>Bug Report Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroticism</td>
<td>1</td>
<td>-.329*</td>
<td></td>
<td></td>
<td></td>
<td>0.034</td>
<td>0.036</td>
<td>-.310*</td>
<td>0.028</td>
<td>0.043</td>
</tr>
<tr>
<td>Extraversion</td>
<td>-.329*</td>
<td>1</td>
<td>-.33**</td>
<td></td>
<td></td>
<td>0.034</td>
<td>0.036</td>
<td>-.241*</td>
<td>0.028</td>
<td>0.043</td>
</tr>
<tr>
<td>Openness to experience</td>
<td>-.33**</td>
<td>-.33**</td>
<td>1</td>
<td></td>
<td></td>
<td>0.034</td>
<td>0.036</td>
<td>-.241*</td>
<td>0.028</td>
<td>0.043</td>
</tr>
<tr>
<td>Agreeableness</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>0.036</td>
<td>0.036</td>
<td>-.241*</td>
<td>0.028</td>
<td>0.043</td>
</tr>
<tr>
<td>Osum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owsum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bug Location Rate</td>
<td>-.310*</td>
<td>-.241*</td>
<td>-.241*</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted Fault Density</td>
<td>0.028</td>
<td>-.241*</td>
<td>-.241*</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bug Report Quality</td>
<td>0.043</td>
<td>0.043</td>
<td>0.043</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

5 Threats to validity

This study was conducted on undergraduate students, most with no professional experience in testing, and with a limited number of participants, at two universities in Victoria, Australia. Replication with different, preferably larger, cohorts would be required to have confidence that our results are generally applicable.

Participants were paid $25 for their participation in the study. This reward might have been their primary motivation for participating and thus distorted their efforts. However, the observed enthusiastic participation of the participants suggests that this was not in fact the case.

One possible threat might be the reliability of our software testing effectiveness test. The bugs injected in the instrument do represent realistic bugs. However, the very small scale of the task is not representative of testing large, real-life software systems. Another threat to the construct validity is that the results may be dependent the domain and programming style used in the software forming the basis of the testing task.

6 Discussion

The purpose of this study was to investigate the effect of different personality traits, as captured by the five-factor model, on effectiveness in software testing, as assessed with a custom software testing task. We recruited nearly 50 students to undertake a detailed NEO test for personality traits and to undertake a software testing task. To find any association we closely examined the Pearson correlations of the five factor personality traits with the overall effectiveness in testing. We also combined several different measures of testing effectiveness in different ways. In contrast to our initial assumption and the anecdotal opinions of testing experts, we did not find strong correlation between the personality factors and the testing effectiveness measures.

A weak negative correlation, however, is seen between extraversion and overall effectiveness, as discussed in Section 4.4. This result differs from the study of Shoaib et al. [40] that concluded extravert people were good at exploratory testing. A negative correlation also exists between extraversion and bug location rate. As suggested by the type descriptions [16], extravert people are assertive, active and talkative. As such it is rather surprising that people who are enthusiastic about communication may not be as good as others at reporting bugs.
Another personality trait, conscientiousness, showed a weak positive correlation with bug location rate and a weak negative correlation with weighted fault density (Section 4.4). Bug location rate takes time spent in testing into account with the total number of bugs found in testing. According to [16] conscientious people are determined, punctual and reliable that complements the positive correlation between bug location and conscientiousness. On the other hand, weighted fault density is more concerned with the severity of found bugs. The negative correlation of conscientiousness with weighted fault density raises the question whether highly conscientious individuals do not pay as much attention to the severity of bugs? The above two findings indicate that highly conscientious individuals might be more interested towards finding more bugs without giving as much attention to the severity of the bugs found.

Conscientiousness is closely linked to the Judging-Perceiving dimension of MBTI [23]. According to Capretz and Ahmed [12, 13], sensing and judging individuals categorized by MBTI would be good at testing. Our finding that conscientiousness has influence on testing effectiveness, thus, partially supports Capretz and Ahmed’s perception. However, the exact effect of conscientiousness should be examined by more specific experiments.

7 Conclusions

In this empirical study we attempted to identify if correlations exist between the five major personality traits, as measured by the NEO inventory, and student performance on a representative software testing task, as measured by three performance metrics. We carried out a study with nearly 50 Computer Science and Software Engineering students, assessing their personality traits and their performance on our testing tasks. Our results indicated weak associations between both extraversion and conscientiousness, and effectiveness in testing. We plan to replicate the experiment with a larger number of participants to test the validity of this finding. We plan to replicate our study with different kinds of testing tasks. Ideally we would also like to replicate our study with actual testing professionals. More precise effects of the extraversion and conscientiousness traits should be examined with a more focused experiment in which the influence of the other independent variables can be kept to a minimum.

References


