An Empirical Study of the Effects of Personality on Software Testing

Tanjila Kanij and John Grundy
Swinburne University of Technology
Melbourne, Australia

Robert Merkel
Monash University
Melbourne, Australia
Outline

- Motivation
- Background
- Experimental design
  - Personality Assessment
  - Testing Task and performance assessment
- Results
- Conclusions
Motivation

- Personality impact has been studied on various aspects of Software Engineering e.g. coding, pair programming, teamwork

- Anecdotally, it has been thought that software testers are more conscientious, neurotic, more open to experience…

- But no one really knows!

- In an earlier study of professional testers opinions, we found mixed views on what impacts testing performance and ability
Background

- Personality – MBTI vs Five Factor model
- Human factor impact on software testing
  - Experience, attitude, organisational impact
- Personality factors and programming
  - Specific MBTI traits over-represented, but...
  - Five factor-based assessment suggested no predictors
- Personality factors and software engineering
  - Capretz’s studies – sensing, thinking, judging, intuitive critical factors (MBTI); Feldt et al’s “clusters” of factors
  - Clark et al’s conscientiousness, introversion findings
“Expert Opinion”

- Armour – “nose for testing”
- Pettichord – tolerate tedium, skeptical, handle conflicts
- Pol et al – creative, accurate, strict in methodology
- Capretz and Ahmed – job responsibility analysis – attention to detail, good organisational skills, sensing and judging
Our Study

- Empirically determine relationship between personality type using Five Factor model and testing performance

- Use Computer Science & Software Engineering students as the population to sample

- Quasi-experiment of:
  - testing task to complete
  - personality assessment
  - performance assessment
Assessment of Personality

- Five Factor model

- NEO PI-3 inventory, measuring:
  - **Extraversion (E):** related to sociability, assertiveness, talkativeness and activeness.
  - **Agreeableness (A):** the expressive quality of admirable human aspects of personality
  - **Conscientiousness (C):** “Will to achieve” - purposeful, strong-willed and determined
  - **Neuroticism (N):** 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.
Testing Task, Performance Assessment

- Test faulty Java program (derived from assignment in another unit)

- 18 de-identified assignments used to craft one with common (and uncommon) faults – 20 in all; 1017 lines code; max method cyclometric complexity of 7

- Classified severity using Hutchison’s taxonomy

- Compared injected bugs to Knuth’s errors and Eisenstadt’s bug war stories
## Faults & Classification

<table>
<thead>
<tr>
<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 on the console</td>
<td>Missing functionality</td>
<td>Students</td>
<td>Low</td>
<td>forgotten function</td>
<td>unsolved</td>
</tr>
<tr>
<td>2</td>
<td>Pair of cities are printed with their names</td>
<td>Pairs are printed with numbers</td>
<td>Wrong implementation</td>
<td>Students</td>
<td>Medium</td>
<td>algorithm awry</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Prints right number of pairs possible among the given cities</td>
<td>Prints wrong number of pair of cities</td>
<td>Wrong implementation</td>
<td>Researchers</td>
<td>High</td>
<td>blunder or botch</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Prints right distance between the pair of cities</td>
<td>Prints wrong distance between pair of cities</td>
<td>Wrong function call</td>
<td>Researchers</td>
<td>High</td>
<td>mismatch between modules</td>
<td>des.Logic</td>
</tr>
<tr>
<td>5</td>
<td>Prints pairs with right (name/number) of cities</td>
<td>Prints pairs with wrong (name/number) of cities</td>
<td></td>
<td>Researchers</td>
<td>High</td>
<td>blunder or botch</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Prints right number of topologies possible among the given cities</td>
<td>Prints wrong number of topologies among the given cities</td>
<td>Wrong index</td>
<td>Students</td>
<td>High</td>
<td>blunder or botch</td>
<td>lex/var</td>
</tr>
<tr>
<td>7</td>
<td>Prints right topology distance</td>
<td>Prints wrong topology distance</td>
<td>Wrong index</td>
<td>Students</td>
<td>High</td>
<td>blunder or botch</td>
<td>lex/var</td>
</tr>
<tr>
<td>8</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>blunder or botch</td>
<td>var</td>
</tr>
<tr>
<td>9</td>
<td>Not printing duplicate topologies</td>
<td>Printing duplicate topology</td>
<td></td>
<td>Researchers</td>
<td>Medium</td>
<td>blunder or botch</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Right display of distance unit</td>
<td>Wrong display of distance unit</td>
<td>Wrong string literal</td>
<td>Researchers</td>
<td>Medium</td>
<td>trivial typo</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Should calculate distance for all values</td>
<td>Cannot calculate topologies for special value (gives exception with 2 cities)</td>
<td></td>
<td>Students</td>
<td>Critical</td>
<td>Forgotten function</td>
<td>lang</td>
</tr>
<tr>
<td>12</td>
<td>Calculates right distance for cities with &lt;0 degree</td>
<td>Calculates wrong distance for cities with &lt;0 degree</td>
<td>Missing Bracket</td>
<td>Students</td>
<td>High</td>
<td>trivial typo</td>
<td>lex</td>
</tr>
<tr>
<td>13</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>Language liability</td>
<td>mem</td>
</tr>
</tbody>
</table>
Assessment Metrics

- Bug location rate (BLR):
  - number bugs found / time taken (mins)

- Weighted fault density (WFD):
  - sum of (weight * severity) / number found

- Bug report quality (BRQ):
  - assessed using the IEEE standard of Test Documentation

- Overall effectiveness
  - Total score (BLR+WFD+BRQ) vs
  - Weighted total score (0.3*BLR+0.3*WFD+0.4*BRQ)
Results

- 48 students; 18-35 years old; 69% male
- 23% had professional experience in testing
- 31% had done specialised testing unit
- 27% had used testing tools
- Shapiro-Wilk Test indicated that our population distributions do not differ significantly from normality, for the NEO personality inventory used to assess personality
## Distribution of Scores

**Table 2: Distribution of scores (N = 48)**

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Average</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroticism (N)</td>
<td>32</td>
<td>74</td>
<td>54.94</td>
<td>10.38</td>
</tr>
<tr>
<td>Extraversion (E)</td>
<td>20</td>
<td>70</td>
<td>50.42</td>
<td>9.32</td>
</tr>
<tr>
<td>Openness to experience (O)</td>
<td>35</td>
<td>80</td>
<td>54.36</td>
<td>10.02</td>
</tr>
<tr>
<td>Agreeableness (A)</td>
<td>29</td>
<td>74</td>
<td>48.27</td>
<td>9.62</td>
</tr>
<tr>
<td>Conscientiousness (C)</td>
<td>27</td>
<td>66</td>
<td>47.25</td>
<td>8.62</td>
</tr>
<tr>
<td>$O_{sum}$</td>
<td>0.63</td>
<td>3.87</td>
<td>1.91</td>
<td>0.85</td>
</tr>
<tr>
<td>$O_{wsum}$</td>
<td>0.24</td>
<td>1.51</td>
<td>0.73</td>
<td>0.34</td>
</tr>
<tr>
<td>Bug Location Rate (BLR)</td>
<td>0.02</td>
<td>0.37</td>
<td>0.12</td>
<td>0.063</td>
</tr>
<tr>
<td>Weighted Fault Density (WFD)</td>
<td>0.1</td>
<td>0.33</td>
<td>0.23</td>
<td>0.07</td>
</tr>
<tr>
<td>Bug Report Quality (BRQ)</td>
<td>0.5</td>
<td>3.5</td>
<td>1.56</td>
<td>0.83</td>
</tr>
</tbody>
</table>
### Personality traits vs testing effectiveness

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

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>E</th>
<th>O</th>
<th>A</th>
<th>C</th>
<th>O_{sum}</th>
<th>O_{wsum}</th>
<th>BLR</th>
<th>WFD</th>
<th>BRQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroticism</td>
<td>1</td>
<td>-0.329</td>
<td>-0.136</td>
<td>-0.135</td>
<td>-0.457</td>
<td>0.034</td>
<td>0.036</td>
<td>0.122</td>
<td>0.02</td>
<td>0.043</td>
</tr>
<tr>
<td>Extraversion</td>
<td>1</td>
<td>0.401</td>
<td>-0.231</td>
<td>0.375</td>
<td>-0.267</td>
<td>-0.267</td>
<td>0.038</td>
<td>-0.133</td>
<td>-0.0.191</td>
<td></td>
</tr>
<tr>
<td>Openness</td>
<td>1</td>
<td>-0.235</td>
<td>0.177</td>
<td>0.161</td>
<td>0.165</td>
<td>-0.025</td>
<td>-0.154</td>
<td>0.179</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness</td>
<td>1</td>
<td>-0.121</td>
<td>0.167</td>
<td>0.173</td>
<td>-0.034</td>
<td>-0.215</td>
<td>0.191</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>1</td>
<td>0.026</td>
<td>0.026</td>
<td>0.251</td>
<td>-0.241</td>
<td>0.028</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_{sum}</td>
<td>1</td>
<td>1.000**-</td>
<td>0.258</td>
<td>0.085</td>
<td>0.996</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O_{wsum}</td>
<td>1</td>
<td>0.248</td>
<td>0.071</td>
<td>0.998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLR</td>
<td>1</td>
<td>1</td>
<td>-0.310</td>
<td>0.214</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WFD</td>
<td>1</td>
<td>1</td>
<td>0.028</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRQ</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outcomes

- Weak negative correlation – extraversion vs overall effectiveness (differs from previous studies)

- Weak negative correlation – extraversion and bug report quality - surprising?

- Weak positive correlation – conscientiousness and bug location rate – expected?

- Weak negative correlation - conscientiousness and weighted fault density – more vs severity (quantity vs quality?!)

Implications

- Who makes a better tester – does personality matter??
- Need to be conscientious 😊
- Extroversion-related qualities might negatively impact bug reporting?!
- Teaching testing – bug location vs bug severity vs bug report quality
  - Not all bugs are equal!
- Assessing testing – when know student / tester has done a good job??
Summary

- Empirical study of CS&SE students to examine impact of personality, as measured by Five Factor model, on testing effectiveness

- Moderate size Java program with 20 errors, ranging in severity, derived from older student exemplars & widely used standard

- Most personality indicators didn’t seem to impact testing effectiveness in our study

- Weak +ve impact of conscientiousness on finding bugs, but –ve on severity – quantity vs quality??

- Weak –ve impact of extraversion on effectiveness
References


