Assessing Incremental Testing Practices and Their Impact on Project Outcomes

Ayaan M. Kazerouni, Clifford A. Shaffer, Stephen H. Edwards, Francisco Servant

Department of Computer Science, Virginia Tech

SIGCSE 2019, Minneapolis, MN
Thursday, February 28, 2018
Study Context

- Third year (post-CS2) Data Structures & Algorithms course
- 157 students
- 4 assignments
  - Median 1.4 kLOC
  - 3-4 weeks long
- 415 implementations (unbalanced)
Contributions

- Family of metrics to assess incremental testing

- Empirical study
  1. How does the balance of testing effort relate to project outcomes?
  2. How does the sequence of testing effort relate to project outcomes?
Better Feedback on Process

Programming effort

Time →

| Work Session | Solution | Tests | Method A | Method B | Method C |

Feedback

Correctness: 100%
Code coverage: 89%
Procrastination: 75% [1]
Balance of testing
Thoroughness of testing

Assessing Incremental Testing
Proposed Metrics of Testing Effort

Synthetic example: sequence of developer activity

<table>
<thead>
<tr>
<th>Solution code</th>
<th>Test code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method A</td>
</tr>
<tr>
<td></td>
<td>Method B</td>
</tr>
<tr>
<td></td>
<td>Method C</td>
</tr>
<tr>
<td></td>
<td>Any method</td>
</tr>
</tbody>
</table>

- Method A
- Method B
- Method C
- Any method
Proposed Metrics of Testing Effort

Synthetic example: sequence of developer activity

Project-wide Overall Testing Effort

\[
\frac{T}{S + T}
\]
Proposed Metrics of Testing Effort

Synthetic example: sequence of developer activity

Project-wide Overall Testing Effort

\[ \frac{T}{S + T} \]

Project-wide per-Session Testing Effort

\[ \text{median} \left\{ \frac{T_s}{S_s + T_s} \right\} \]
Proposed Metrics of Testing Effort

Synthetic example: sequence of developer activity

Project-wide Overall Testing Effort

Project-wide per-Session Testing Effort

Method-specific Overall Testing Effort

Solution code  Test code

- Method A
- Method B
- Method C
- Any method

\[
\text{median} \left\{ \frac{T_{s}}{S_s + T_{s}} \right\}
\]

\[
\text{median} \left\{ \frac{T_{m}}{S_m + T_{m}} \right\}
\]
Proposed Metrics of Testing Effort

Synthetic example: sequence of developer activity

![Sequence of developer activity]

**Project-wide Overall Testing Effort**

\[
\frac{T}{S + T}
\]

**Project-wide per-Session Testing Effort**

\[
\text{median}\left\{\frac{T_s}{S_s + T_s}\right\}
\]

**Method-specific Overall Testing Effort**

\[
\text{median}\left\{\frac{T_m}{S_m + T_m}\right\}
\]

**Method-specific per-Session Testing Effort**

\[
\text{median}\left\{\frac{T_{ms}}{S_{ms} + T_{ms}}\right\}_{mm}
\]
Proposed Metrics of Testing Effort

Synthetic example: sequence of developer activity

Project-wide Overall Testing Effort

Project-wide per-Session Testing Effort

Method-specific Overall Testing Effort

Method-specific per-Session Testing Effort

Method-specific Overall Sequence of Testing Effort

Solution code

Test code

Method A

Method B

Method C

Any method

= Method is “finalised”
Motivating Example from Fall 2016

Fig. 1: Good Test Writing Process

Fig. 2: Poor Test Writing Process
Empirical Study
Data Collection

- 400+ project implementations

**Edit Event**

Type: Edit
Time: 1477672862
Snapshot Id: 23479b3

<table>
<thead>
<tr>
<th>Type</th>
<th>Size</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change in method <code>insertFront</code></td>
<td>+5</td>
<td>12:41:02</td>
</tr>
<tr>
<td>Change in method <code>getSize</code></td>
<td>+1</td>
<td>12:41:02</td>
</tr>
<tr>
<td>Change in test for <code>insertFront</code></td>
<td>+3</td>
<td>12:41:02</td>
</tr>
</tbody>
</table>
Study Design

- Fixed effects: 5 measures of testing effort
- Random effects: students, assignments
- Outcome variables:
  - Correctness, measured by the percentage of reference tests passed
  - Code coverage achieved by the student’s own test suite

Mixed effects model: repeated measures for each student, and for each assignment.
Results
Project-wide Overall Testing Effort

<table>
<thead>
<tr>
<th>Solution Code Effort</th>
<th>Test Code Effort</th>
</tr>
</thead>
</table>

Expectation: Positive relationship with correctness and code coverage.

<table>
<thead>
<tr>
<th>Correctness</th>
<th>Code Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression estimate</td>
<td>Regression estimate</td>
</tr>
<tr>
<td>$0.30$</td>
<td>$0.23$</td>
</tr>
<tr>
<td>$&lt; 0.001 \star$</td>
<td>$&lt; 0.001 \star$</td>
</tr>
</tbody>
</table>

- Implementations with a higher project-wide testing effort achieved:
  - Higher semantic correctness
  - Higher code coverage
Project-wide per-Session Testing Effort

Implementations with higher testing effort within each work session achieved
- Higher semantic correctness
- Higher code coverage

Expectation: Positive relationship with correctness and code coverage.

<table>
<thead>
<tr>
<th>Correctness</th>
<th>Code Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression estimate</td>
<td>$p$</td>
</tr>
<tr>
<td>0.30</td>
<td>0.005 $^*$</td>
</tr>
</tbody>
</table>

• Implementations with higher testing effort within each work session achieved
  • Higher semantic correctness
  • Higher code coverage
Motivating Example (Reprise)

Fig. 1: Good Test Writing Process

Fig. 2: Poor Test Writing Process
Method-specific Sequence of Testing Effort

Expectation: Positive or no relationship with project outcomes.

<table>
<thead>
<tr>
<th>Correctness</th>
<th>Code Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression estimate</td>
<td>p</td>
</tr>
<tr>
<td>--</td>
<td>0.10</td>
</tr>
</tbody>
</table>

- Implementations where a higher proportion of testing for a method was done before the method was finalised, achieved:
  - No significant change in correctness
  - Lower code coverage
Putting It All Together

1

+ Correctness
+ Code
+ Coverage

2

+ Correctness
+ Code
+ Coverage

3

? Correctness
- Code Coverage
Closing Remarks
Summary

- Quantified test-writing practices
- Empirical study
  - Higher testing effort is good (whole project and per-method)
  - Higher testing effort per work session is good
  - No such relationship on a per-method basis
  - Higher testing effort before finalizing relevant solution code
    - Does not lead to improved correctness
    - Negative relationship with code coverage

- Next step: Design and deploy automated interventions for continuous feedback

ayaan@vt.edu    https://github.com/ayaankazerouni/incremental-testing
Bonus Material
Per-session Testing Effort: Distribution
Method-specific Sequence of Testing Effort

Method-specific Overall Sequence of Testing Effort (MOS)

Number of Observations

0.0 0.2 0.4 0.6 0.8 1.0

0 10 20 30 40
Method-specific per-Session Testing Effort

<table>
<thead>
<tr>
<th>Work Session</th>
<th>Method A</th>
<th>Method B</th>
<th>Method C</th>
</tr>
</thead>
</table>

Expectation: Positive relationship with both project outcomes.

<table>
<thead>
<tr>
<th>Correctness</th>
<th>Condition Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression estimate</td>
<td>$p$</td>
</tr>
<tr>
<td>--</td>
<td>0.10</td>
</tr>
</tbody>
</table>

• Higher testing effort per-method, per-session achieved:
  • Higher condition coverage
  • No significant change in
## Mixed effects model (Process)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Correctness</th>
<th></th>
<th>Code Coverage</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regression estimate</td>
<td>p</td>
<td>Regression estimate</td>
<td>p</td>
</tr>
<tr>
<td>Testing per-Session</td>
<td>0.30</td>
<td>0.005 *</td>
<td>0.12</td>
<td>0.008 *</td>
</tr>
<tr>
<td>Testing per-Session per-Method</td>
<td>--</td>
<td>0.10</td>
<td>0.09</td>
<td>0.002 *</td>
</tr>
<tr>
<td>Sequence of testing</td>
<td>--</td>
<td>0.62</td>
<td>-0.06</td>
<td>0.02 *</td>
</tr>
</tbody>
</table>
## Mixed effects model (Overall)

<table>
<thead>
<tr>
<th>Metric</th>
<th>Correctness</th>
<th>Code Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Regression estimate</td>
<td>p</td>
</tr>
<tr>
<td><strong>Testing</strong></td>
<td>0.30</td>
<td>&lt; 0.001 *</td>
</tr>
<tr>
<td><strong>Testing per-Method</strong></td>
<td>--</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>Testing per-Session</strong></td>
<td>--</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>Testing per-Session, per-Method</strong></td>
<td>--</td>
<td>0.97</td>
</tr>
<tr>
<td><strong>Sequence of testing</strong></td>
<td>--</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Fixed effects $R^2 = 5\%$

Fixed effects $R^2 = 10\%$