Discussion of Class Experiment

CMSC 735
Jeff Carver
Dec. 10, 2001

Experimental Setting

• \textit{In vitro}: in the laboratory under controlled conditions
  – \textit{In vivo} would mean that it was done in a real work environment
  – Classroom setting imposes its own constraints:
    • Examples??
      – No control group
      – No control over subject population
Classifying the Experiment

• Types of analysis
  – Qualitative analysis (mostly): naturalistic observation, discovery oriented
    • Not a lot of hypotheses; want to be able to propose well-founded ones
  – Also some quantitative analysis (# of defects)

• Levels of variable relationship
  – Descriptive (most likely)
  – Correlative (desired)

Experimental Goals

• GQM
  – To analyze:
    • PBR for the purpose of understanding with respect to learning from the viewpoint of the researcher.
    • PBR for the purpose of understanding and improving with respect to feasibility from the viewpoint of the researcher.
Experimental Goals

• Pedagogical:
  – To introduce software experimentation
  – To demonstrate the difficulties of thinking about software processes, and give students experience with a strategy – Observation
  – To familiarize students with the idea of evolving techniques based on feedback

Experimental Design

<table>
<thead>
<tr>
<th>Group A (7 teams)</th>
<th>Group B (6 teams)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment 1</td>
<td>Treatment 2</td>
</tr>
<tr>
<td>PBR Requirements Inspection</td>
<td>PBR Requirements Inspection</td>
</tr>
<tr>
<td>Loan Arranger</td>
<td>Parking Garage</td>
</tr>
<tr>
<td>(User Perspective)</td>
<td>(User Perspective)</td>
</tr>
<tr>
<td>Parking Garage</td>
<td>Loan Arranger</td>
</tr>
<tr>
<td>(User Perspective)</td>
<td>(User Perspective)</td>
</tr>
</tbody>
</table>

• Independent Variables:
  – Previous Domain Knowledge
  – Team Composition
    • Knowledge of English
    • Industry Experience
Experimental Artifacts

<table>
<thead>
<tr>
<th></th>
<th>Parking Garage</th>
<th>Loan Arranger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain Experience</td>
<td>Familiar</td>
<td>Unfamiliar</td>
</tr>
<tr>
<td>Pages</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Requirements</td>
<td>21 Functional</td>
<td>26 Functional</td>
</tr>
<tr>
<td></td>
<td>9 Nonfunctional</td>
<td>4 Nonfunctional</td>
</tr>
<tr>
<td>Seeded Defects</td>
<td>32</td>
<td>18</td>
</tr>
</tbody>
</table>

GQM: Questions & Metrics

- Does having the process observer/guide affect the results of the process?
  - Measure the percentage of (known) defects the (6) teams reviewing the Parking Garage requirements and the (6) teams reviewing the Loan Arranger requirements report, compared with historical data for those requirements
  - Qualitative data collected in the reports
Experimental Results

• Does having the process/observer guide affect the results of the process?
  – 3 teams said that they would rather work as a team instead of Executor and Observer

GQM: Questions & Metrics

• Does PBR work more or less effectively when the reviewer has some experience with the problem domain?
  – Number of defects found by inspectors in each problem domain
  – Qualitative data from reports
Experimental Results

• Does PBR work more or less effectively when the reviewer has some experience with the problem domain?
  – Quantitative
    • PG Average = 14.2% of known defects found
    • LA Average = 17.9% of known defects found
    • T-test p-value (two-tail) = .33
      – No statistical significance
  – Qualitative
    • 4 teams indicated that they believed domain knowledge was or would have been helpful

GQM: Questions & Metrics

• What can we do to improve PBR?
  – Qualitative data from reports
Experimental Results

• What can we do to improve PBR?
  – 6 teams said that the questions could be improved
    • Adding more specific kinds
      – Domain specific
      – Organization specific
  – 4 teams wanted to reorder the steps
    • Most want to find the system functionality either before finding Use Cases or before uncovering defects
  – 3 teams were unclear about which “user” they were to take the perspective of

GQM: Questions & Metrics

• What effect does using a technique multiple times have on the executor’s performance?
  – Compare effectiveness of the second set of inspectors to that of the first set
  – Qualitative data in reports
Experimental Results

• What effect does using a technique multiple times have on the executor’s performance?
  – Quantitative
    • PG Requirements only
      – First Inspection 17.7% of known defects found
      – Second Inspection 14.1% of known defects found
      – T-test (two-tailed) p-value = .25 (In the wrong direction)
    • LA Requirements only
      – First Inspection 15.1% of known defects found
      – Second Inspection 21.3% of known defects found
      – T-test (one-tail) p-value = .15

• What effect does using a technique multiple times have on the executor’s performance?
  – Qualitative
    • 3 teams said they understood the steps better
    • 7 teams said they were able to reorder or curtail steps
    • 5 teams said they were more confident and efficient
    • 3 teams said they knew better what to tell the observer
GQM: Questions & Metrics

• How much improvement would we get by doing a “team inspection”?
  – Look at number of defects found by Observer that were missed by Executor

Experimental Results

• How much improvement would we get by doing a “team inspection”?

<table>
<thead>
<tr>
<th></th>
<th>PG Average</th>
<th>LA Average</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without Observer</td>
<td>14.2%</td>
<td>17.9%</td>
<td>33</td>
</tr>
<tr>
<td>With Observer</td>
<td>16.8%</td>
<td>18.3%</td>
<td>66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PG</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Defects found by observers</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>
GQM: Questions & Metrics

• Does Thinking-aloud have any affect on the inspection process?
  – Qualitative data from reports
Experimental Results

• Does Thinking-aloud have any affect on the inspection process?
  – Comments, questions?

Threats to Validity

• Internal:
  - History:
    • Results of later treatments may be attributed to events that occurred between treatments
  - Maturation:
    • Processes occurring within subjects may change over time
  - Testing:
    • Results may vary over time as subjects get more comfortable with testing procedures
  - Instrumentation:
    • Results may differ with different measures
  - Selection:
    • Results may differ because of the type of subjects in different groups
  - Process Conformance:
    • Results may differ because procedure was not followed
Threats to Validity

• Internal: potential problems in interpretation of data from this experiment
  – Not all questions can be addressed by this experiment because certain variables are *confounded* between groups, i.e. background and experience of inspectors
  – Small number of subjects leaves two options
    • Make large groups and confound variables
    • Use all combinations of variables but have small groups

• External:
  – Are results valid outside of this class?
    • More experienced subjects?
    • Less experienced subjects?
    • Professional developers?
  – Are results valid for other requirements documents?
    • Different formats?
    • Different domains?
    • Different Languages?