Statistical Tune Up of the Peer Review Engine

Tom Lienhard
Thomas_G_Lienhard@Raytheon.com

November, 2009
Overview

Project Definition:
Reduce Rework by Reducing Defect Leakage

Currently, over 30% of the Software Engineering effort is consumed reworking products already deemed “fit-for-purpose”. A major contributor to this is defect leakage. Defect leakage is calculated as a percentage by summing the defects attributable to a specific phase that are detected in later phases divided by the total number of defects attributable to that phase. Defect leakage is a good indicator of the quality of the different phases of the software process. Defect leakage for some software development phases is as high as 75%, whereas our goal is set at 20%. Not catching and correcting defects at the earliest point in the process leads to cost and budget over-runs due to excessive rework. By investigating what types of defects go undetected during the various phases, corrections can be introduced into the process to help identify the top defect types.
Agenda – Six Sigma Toolbox Examples

Thought Process Map
Process Map
Failure Mode and Effect Analysis
What Was Learned So Far
Product Scorecard
What Was Learned So Far – Part 2
Improvement Goal
Distributional Characterization of Data
What Was Learned So Far – Part 3
DOE Conducted
What Was Learned So Far – Part 4
Results
Conclusion
**Thought Process Map – Where Are We Headed?**

- **Reduce SW Defect Leakage**
  - What is a defect?
  - What is leakage? (Use definitions from Scorecard)
  - How to reduce SW Defect Leakage?

- **Minimize Defects from entering into the SW development process**
  - What is the process?
  - Is there one?
  - SW4205760, Rev C
  - How to improve the in-phase defect detection process?

- **Improve the in-phase defect detection process**
  - How do we measure the process?
  - Gather data
  - What data?

- **Current process**
  - Process Map
  - What are the possible weaknesses?
  - FMEA/CE
  - Postmortems

- **What is the current leakage?**
  - Gather data
  - What data?
  - My Green Belt
  - Other Black Belts

- **What is the current leakage?**
  - Gathers data
  - What data?

- **What is the leakage?**
  - Data from Program Tracking System
  - Organize data to effectively analyze
  - What data?

- **Effort associated for each cell of worksheet**
  - Use Industry Numbers for effort by phase
  - % Leakage by Phase/Total effort by Phase/Total
  - $ by Phase/Tota

- **Use Industry Numbers for effort by phase (Bob Rova - Motorola, Ti, Hughes)**
  - (Industry Costs.xls)
  - Where introduced
  - Where found
  - (Org/RJ/NGC Scorecard.xls)

- **What about older projects?**
  - No data available unless using SSDP Rev C
  - Data from Program Tracking System

- **Organize data to effectively analyze**
  - Defect #s, Types, Phase intro/Detected
  - Defined in SSDP Rev C
  - X Bar Range and Pareto.xls
  - BARRIER-

- **What data?**
  - What data? BARRIER-
  - Lack of data?

- **Data from Program Tracking System**
  - Data for FMEA
  - What's important to "customer"
  - What have we learned?
  - Industry Costs.xls

- **What is data showing us?**
  - Identify and Remove Special Causes
  - What is the data any good?
  - Validate Measurement System
  - MSE(KAPPA/ICC or Nested Design)
  - Not Adequate
  - ICC .86
  - Adequate
  - Common checklists
  - SQA Process Evals.
  - Updated training
  - Requirement people
  - Roles/respons.
  - Process knowledge
  - FMEA.xls

- **Characterize**
  - What are important factors?
  - Drill down into the data/NEM/Control Charts
  - X Bar R Chart shows prediction range by phase
  - Shows variation within/between phases
  - Pareto charts show defect types, where intro, found
  - (X Bar Range and Pareto.xls)

- **Stable process**
  - C Chart for process C Chart for phases
  - How good classification?
  - How good categorize?

- **Remove Common cause**
  - No
  - Yes
  - Drill down into the data/NEM/Control Charts
  - Characterize Optimize Process
  - Choose of factors

- **Make and communicate improvements**
  - Newsletters
  - Liaison Meetings
  - Common checklists, # of people, Moderator

- **Set up control plan and Use Control Charts to Monitor Effectiveness of Improvements**
  - Did change cause improvement?
  - Did change cause improvement?
  - Improvement

- **Update process to ensure data has higher confidence rate/ train**
  - Updated training, Common checklists, # of people, Moderator
Process Map – Walk the Process

Draft Product Review Team Software Plans

In-Phase Peer Review Process

"Fit-for-Purpose" Product Review Records

Peer Review Planning

Review team - with assigned roles and responsibilities
Evaluation criteria
Notice and agenda

Product Review

S - Product for review
C - Evaluation criteria
C - Reviewers roles and responsibilities
C - Notice and agenda

Review Meeting

S - Product for review
C - Reviewers reviewed the product
C - Review has adequate representation
N - Proposed defects

Proposed defects
Time spent reviewing

Agree on and document defects
Capture review meeting metrics
Gain consensus on review outcome

Rework Product or Disposition Defects?

Rework

S - Product for review
S - Agreed to defects
C - Reviewers

Verify Disposition of Agreed to Defects

Dispositioned?

Yes

No

Fit-for-Purpose product
Peer review records

Corrected product
Defect disposition
Time spent correcting

x - Critical (statistical proven critical)
n - Noise (can’t or choose not to control)
s - SOP (the standard way to do it)
c - Controllable (can be changed to see effect)
## Failure Mode and Effect Analysis – How Can We Mess This Up?

### Process or Product Name: In-Phase Peer Review

### Responsible: Tucson Software Engineering Process Group

### Prepared by: Tom Lienhard & Team

### FMEA Date (Orig) 23 Sept 99 (Rev) 29 Sept 99

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the process step/input under investigation?</td>
<td>In what ways does the process step go wrong?</td>
<td>What is the impact on the Key Output Variables (Customer Requirements) or internal requirements?</td>
<td>Severity 9 - Defects go to customer</td>
<td>Severity 6 - Defects cause rework</td>
<td>Severity 3 - Data not collected</td>
<td>Severity 1 - No harm/no foul</td>
<td>Occurrence 9 - Regular occurrence</td>
<td>Occurrence 6 - Occurs more than occasionally</td>
<td>Occurrence 3 - Occurs occasionally</td>
<td>Occurrence 1 - Rare occurrence</td>
</tr>
<tr>
<td>Peer Review Planning</td>
<td>No Review Team identified upfront w/ review package/roles &amp; responsibilities</td>
<td>Product not reviewed by appropriate disciplines</td>
<td>6</td>
<td>Lack of process awareness</td>
<td>9</td>
<td>SW plans do not require this</td>
<td>SEPG/SQA and peer review of plans</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>No product evaluation criteria identified w/ review package</td>
<td>Product not reviewed to customer and/or process requirements</td>
<td>6</td>
<td>Lack of process awareness</td>
<td>9</td>
<td>SW plans do not require this</td>
<td>SEPG/SQA and peer review of plans</td>
<td>1</td>
<td>9</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>No notice or agenda with review package</td>
<td>Team not able to give adequate review time</td>
<td>6</td>
<td>Lack of process awareness</td>
<td>9</td>
<td>SW plans do not require this</td>
<td>SEPG/SQA and peer review of plans</td>
<td>1</td>
<td>6</td>
<td>216</td>
</tr>
<tr>
<td>Product Review</td>
<td>Product not reviewed</td>
<td>Defects not found</td>
<td>9</td>
<td>Adequate time not given to review</td>
<td>6</td>
<td>SW plans do not require this</td>
<td>SEPG/SQA and peer review of plans</td>
<td>1</td>
<td>3</td>
<td>162</td>
</tr>
<tr>
<td></td>
<td>No or inadequate evaluation criteria</td>
<td>9</td>
<td>SW plans do not require this</td>
<td>SEPG/SQA and peer review of plans</td>
<td>1</td>
<td>3</td>
<td>None - cultural thing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inappropriate reviewers</td>
<td>9</td>
<td>SW plans do not require this</td>
<td>SEPG/SQA and peer review of plans</td>
<td>1</td>
<td>3</td>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metrics not captured</td>
<td>Organization quantitative data incorrect/incomplete</td>
<td>3</td>
<td>SW plans do not require this</td>
<td>SEPG/SQA and peer review of plans</td>
<td>1</td>
<td>3</td>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lack of process awareness</td>
<td>9</td>
<td>SW plans do not require this</td>
<td>SEPG/SQA and peer review of plans</td>
<td>9</td>
<td>486</td>
<td>Create generic checklists for site (to highest level) and conduct process evaluations</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Metrics not captured

- Organization quantitative data incorrect/incomplete
- SW plans do not require this
- Lack of process awareness

### Actions Recommended

- Re-train moderator and conduct process evaluations
Potential causes (factors) high RPN which kept showing up over and over on the FMEA:

- Inappropriate review team (“wrong” moderator, dominant, inexperienced, or yes-people made up the team)
- Lack of process awareness (both unintentional and deliberate)
- No or inadequate review criteria (review what is there not what is missing, biased review based on experience with phase)

Plan to minimize the occurrence and increase the detection:

- Update the process to highlight required participants, their roles and responsibilities on the Notice and Agenda
- Roll-out Peer Review training
- Have SQA perform peer review process evaluations
- Generate common evaluation criteria for all software products that can be used across the entire organization

Use what was learned about factors as an input into DOE
### Product Scorecard

Number of defects identified by phase introduced and phase detected (Modified Software Worksheet from Product Scorecard)

<table>
<thead>
<tr>
<th>Phase Detected</th>
<th>Planning</th>
<th>Customer</th>
<th>Rqmts. Analysis</th>
<th>Design</th>
<th>Implementation</th>
<th>Test</th>
<th>Formal Test</th>
<th>Customer Before</th>
<th>TOTAL</th>
<th>Leaked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>Customer</td>
<td>0</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>Rqmts. Analysis</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>14</td>
<td>29</td>
<td>26</td>
<td>71</td>
<td>1</td>
<td>202</td>
<td>141</td>
</tr>
<tr>
<td>Design</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>323</td>
<td>82</td>
<td>29</td>
<td>38</td>
<td>2</td>
<td>475</td>
<td>151</td>
</tr>
<tr>
<td>Implementation</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>220</td>
<td>43</td>
<td>44</td>
<td>10</td>
<td>323</td>
<td>97</td>
</tr>
<tr>
<td>Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>249</td>
<td>30</td>
<td>0</td>
<td>282</td>
<td>30</td>
</tr>
<tr>
<td>Formal Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>597</td>
<td>0</td>
<td>610</td>
<td>0</td>
</tr>
<tr>
<td>Customer Before</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>29</td>
<td>12</td>
<td>65</td>
<td>344</td>
<td>334</td>
<td>360</td>
<td>786</td>
<td>17</td>
<td>1947</td>
<td>428</td>
</tr>
</tbody>
</table>
### Not All Defects Are Created Equal

**Industry Standard** Cost to Detect and Correct Defects (in days)

<table>
<thead>
<tr>
<th>Phase Detected</th>
<th>Planning</th>
<th>Customer</th>
<th>Rqmts. Analysis</th>
<th>Design</th>
<th>Implementation</th>
<th>Test</th>
<th>Formal Test</th>
<th>Customer Before</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>0.07</td>
<td>0.1</td>
<td>0.19</td>
<td>0.17</td>
<td>0.67</td>
<td>1.23</td>
<td>1.4</td>
<td>0.54</td>
</tr>
<tr>
<td>Customer</td>
<td>0.02</td>
<td>0.15</td>
<td>0.7</td>
<td>1</td>
<td>2.03</td>
<td>2.7</td>
<td>3.23</td>
<td>3.07</td>
</tr>
<tr>
<td>Rqmts. Analysis</td>
<td>0</td>
<td>0.1</td>
<td>0.12</td>
<td>0.86</td>
<td>1.13</td>
<td>1.6</td>
<td>0.79</td>
<td>1.65</td>
</tr>
<tr>
<td>Design</td>
<td>0</td>
<td>0.11</td>
<td>0.13</td>
<td>0.13</td>
<td>0.1</td>
<td>0.8</td>
<td>3.13</td>
<td>2.64</td>
</tr>
<tr>
<td>Implementation</td>
<td>0</td>
<td>0</td>
<td>0.17</td>
<td>0.1</td>
<td>0.7</td>
<td>2.1</td>
<td>2.02</td>
<td>2.33</td>
</tr>
<tr>
<td>Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0.03</td>
<td>0.08</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>Formal Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.18</td>
<td>0.25</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>Customer Before</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.7</td>
<td>3.4</td>
</tr>
</tbody>
</table>

*Motorola, Texas Instruments, Hughes Software Implementation of Six Sigma*
## Cost of Rework Due to Defects (in Days)

### Phase Introduced

<table>
<thead>
<tr>
<th>Phase Detected</th>
<th>Planning</th>
<th>Customer</th>
<th>Rqmts. Analysis</th>
<th>Design</th>
<th>Implementation</th>
<th>Test</th>
<th>Formal Test</th>
<th>Customer Before</th>
<th>TOTAL</th>
<th>Leaked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>2.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.03</td>
<td>0</td>
</tr>
<tr>
<td>Customer</td>
<td>0</td>
<td>1.8</td>
<td>1.4</td>
<td>0</td>
<td>4.06</td>
<td>0</td>
<td>16.15</td>
<td>0</td>
<td>23.41</td>
<td>21.61</td>
</tr>
<tr>
<td>Rqmts. Analysis</td>
<td>0</td>
<td>0</td>
<td>7.32</td>
<td>12.04</td>
<td>32.77</td>
<td>41.6</td>
<td>56.09</td>
<td>0.79</td>
<td>150.61</td>
<td>143.29</td>
</tr>
<tr>
<td>Design</td>
<td>0</td>
<td>0</td>
<td>0.13</td>
<td>41.99</td>
<td>8.2</td>
<td>23.2</td>
<td>118.94</td>
<td>5.28</td>
<td>197.74</td>
<td>155.75</td>
</tr>
<tr>
<td>Implementation</td>
<td>0</td>
<td>0</td>
<td>0.17</td>
<td>0.5</td>
<td>154</td>
<td>90.3</td>
<td>88.88</td>
<td>23.3</td>
<td>357.15</td>
<td>203.15</td>
</tr>
<tr>
<td>Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.16</td>
<td>0.03</td>
<td>19.92</td>
<td>4.5</td>
<td>0</td>
<td>24.61</td>
<td>4.69</td>
</tr>
<tr>
<td>Formal Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.34</td>
<td>149.25</td>
<td>0</td>
<td>151.59</td>
<td>2.34</td>
</tr>
<tr>
<td>Customer Before</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.7</td>
<td>13.6</td>
<td>16.3</td>
<td>13.6</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>2.03</strong></td>
<td><strong>1.8</strong></td>
<td><strong>9.02</strong></td>
<td><strong>54.69</strong></td>
<td><strong>199.06</strong></td>
<td><strong>177.36</strong></td>
<td><strong>436.51</strong></td>
<td><strong>42.97</strong></td>
<td><strong>923.44</strong></td>
<td><strong>544.43</strong></td>
</tr>
</tbody>
</table>
What Was Learned, ….. So Far (Part 2)

Using Pareto Charts We Know Where Defects Enter the Process . . .

And Where Those Defects Are Detected by the Process . . .

And Even Which Phases Are Hitting the Bottom Line

What We Learned……..

• Formal test introduced > 30% of defects
• Finding 58% of defects when product is done (i.e., testing)
• 3 Phases account for > 92% of rework due to leakage
Improvement Goal

Impact on bottom line

Baseline defect cost profile

Goal defect cost profile

Same number of total defects introduced in the same phases
Distributional Characteristics of Data

Stable or Unstable?
Is the variation between or within subgroups?

These 3 phases account for > 92% of rework
Stability of Entire Process...

P Chart for In Phase

P = 0.7678

3.0SL = 1.000
P = 0.7678
-3.0SL = 0.00E+00
....Then applied to Each Phase

Looks pretty stable within subgroups (projects)...

Predictable to be between 0% - 100%
Customer process, team spun off to work this
How About Between the Subgroups?

Displays averages “between” the subgroups

*Out-of-control range indicates special cause affecting customers’ process

Displays range “within” the subgroups
ANOVA Confirms Our Suspicions

Analysis of Variance for percent

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>phase</td>
<td>7</td>
<td>10841.9583</td>
<td>1548.8512</td>
<td>8.701</td>
<td>0.000</td>
</tr>
<tr>
<td>project</td>
<td>16</td>
<td>2848.0000</td>
<td>178.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>13689.9583</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variance Components

<table>
<thead>
<tr>
<th>Source</th>
<th>Var Comp.</th>
<th>% of Total</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>phase</td>
<td>456.950</td>
<td>71.97</td>
<td>21.376</td>
</tr>
<tr>
<td>project</td>
<td>178.000</td>
<td>28.03</td>
<td>13.342</td>
</tr>
<tr>
<td>Total</td>
<td>634.950</td>
<td></td>
<td>25.198</td>
</tr>
</tbody>
</table>

Project variation    Phase variation

But we sampled only 3 projects - what does the population look like?
### Descriptive Statistics

#### Variable: Cust1
- **Anderson-Darling Normality Test**
  - A-Squared: 0.446
  - P-Value: 0.081
- **Mean**: 70.8667
- **SD**: 25.4231
- **Variance**: 646.333
- **Skewness**: 1.72300
- **Kurtosis**: 3
- **Minimum**: 55.0000
- **1st Quartile**: 55.0000
- **Median**: 68.0000
- **3rd Quartile**: 73.0000
- **Maximum**: 93.0000
- **95% Confidence Interval for Mu**: 65.0000 to 73.0000
- **95% Confidence Interval for Median**: 65.0000 to 73.0000

#### Variable: Req
- **Anderson-Darling Normality Test**
  - A-Squared: 0.338
  - P-Value: 0.200
- **Mean**: 28.6667
- **SD**: 3.2146
- **Variance**: 10.3333
- **Skewness**: -1.54539
- **Kurtosis**: 3
- **Minimum**: 25.0000
- **1st Quartile**: 25.0000
- **Median**: 30.0000
- **3rd Quartile**: 33.0000
- **Maximum**: 36.6521
- **95% Confidence Interval for Mu**: 25.0000 to 31.0000
- **95% Confidence Interval for Median**: 25.0000 to 31.0000

#### Variable: Design
- **Anderson-Darling Normality Test**
  - A-Squared: 0.259
  - P-Value: 0.384
- **Mean**: 42.2518
- **SD**: 4.8377
- **Variance**: 55.0000
- **Skewness**: 0.259
- **Kurtosis**: 3
- **Minimum**: 55.0000
- **1st Quartile**: 55.0000
- **Median**: 68.0000
- **3rd Quartile**: 73.0000
- **Maximum**: 88.4149
- **95% Confidence Interval for Mu**: 65.0000 to 73.0000
- **95% Confidence Interval for Median**: 65.0000 to 73.0000

#### Variable: Implem
- **Anderson-Darling Normality Test**
  - A-Squared: 0.212
  - P-Value: 0.536
- **Mean**: 58.6271
- **SD**: 2.1042
- **Variance**: 65.0000
- **Skewness**: 0.212
- **Kurtosis**: 3
- **Minimum**: 65.0000
- **1st Quartile**: 65.0000
- **Median**: 68.0000
- **3rd Quartile**: 73.0000
- **Maximum**: 78.7062
- **95% Confidence Interval for Mu**: 65.0000 to 73.0000
- **95% Confidence Interval for Median**: 65.0000 to 73.0000

#### Variable: test
- **Anderson-Darling Normality Test**
  - A-Squared: 0.193
  - P-Value: 0.616
- **Mean**: 73.985
- **SD**: 2.868
- **Variance**: 82.0000
- **Skewness**: 0.193
- **Kurtosis**: 3
- **Minimum**: 82.0000
- **1st Quartile**: 82.0000
- **Median**: 88.0000
- **3rd Quartile**: 93.0000
- **Maximum**: 101.348
- **95% Confidence Interval for Mu**: 82.0000 to 93.0000
- **95% Confidence Interval for Median**: 82.0000 to 93.0000

#### Variable: Form Tst
- **Anderson-Darling Normality Test**
  - A-Squared: 0.189
  - P-Value: 0.631
- **Mean**: 95.516
- **SD**: 0.521
- **Variance**: 97.000
- **Skewness**: 0.189
- **Kurtosis**: 3
- **Minimum**: 95.516
- **1st Quartile**: 95.516
- **Median**: 98.000
- **3rd Quartile**: 99.000
- **Maximum**: 100.484
- **95% Confidence Interval for Mu**: 98.000 to 99.000
- **95% Confidence Interval for Median**: 98.000 to 99.000
If I were a betting man….
the true population means
are within these intervals.

...Gives Us Plausible Population Range
What Was Learned, ..... So Far (Part 3)

- Variation between the phases (72%) is greater than variation between projects (28%)
  - need to work largest source of variation - what changed between, what didn’t, etc.
- If no action is taken 95% confident that
  - the Requirements Phase will find between 21% - 37% of defects in phase
  - the Design Phase will find between 42% - 88% of defects in phase
  - the Implementation Phase will find 59% - 78% of defects in phase

Green - customer process
Red - < 92.5 of rework
Blue - testing process
Design of Experiments (DOE)

Review Ada packages and C++ objects

Four Factors

- experience: (<2 yrs, >2 yrs)
- training: (No, Yes)
- review criteria: (None, Checklist)
- number of reviewers: (2, >2)

Block by Program

- language, management style, schedule pressures

Sixteen Runs $2^{5-1}$ Half fraction

Resolution V Mains compounded w/4ways, 2 w/3ways

Response Variable

- percentage of defects which match SEPG and project leads’

Limitations:

- “chunks of code” reviewed were different
- restrictions on randomization
- hard to find “team” fulfilling factor levels
## DOE Run Results

<table>
<thead>
<tr>
<th>StdOrder</th>
<th>RunOrder</th>
<th>Program</th>
<th>Experience</th>
<th>Training</th>
<th>Criteria</th>
<th>Num People</th>
<th>% Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>70</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>65</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>45</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>55</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>80</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>55</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>70</td>
</tr>
</tbody>
</table>
DOE Run Chart

Run Chart for % Match

% Match

Observation
## Sorting by Response

<table>
<thead>
<tr>
<th>StdOrder</th>
<th>RunOrder</th>
<th>Program</th>
<th>Experience</th>
<th>Training</th>
<th>Criteria</th>
<th>Num People</th>
<th>% Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>80</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>70</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>70</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>65</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>55</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>55</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>45</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>30</td>
</tr>
</tbody>
</table>

*Might have something here*
Normal Probability Plot of the Effects
(response is % Match, Alpha = .10)

Data looks pretty normal

Shows Training, Criteria, Experience as the influential factors
Fractional Factorial Fit
Estimated Effects and Coefficients for % (coded units)

<table>
<thead>
<tr>
<th>Term</th>
<th>Effect</th>
<th>Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td></td>
<td>60.312</td>
</tr>
<tr>
<td>Program</td>
<td>-1.875</td>
<td>-0.938</td>
</tr>
<tr>
<td>Experien</td>
<td>13.125</td>
<td>6.562</td>
</tr>
<tr>
<td>Training</td>
<td>19.375</td>
<td>9.687</td>
</tr>
<tr>
<td>Criteria</td>
<td>-13.125</td>
<td>-6.562</td>
</tr>
<tr>
<td>Num Peop</td>
<td>3.125</td>
<td>1.562</td>
</tr>
<tr>
<td>Program*Experien</td>
<td>-1.875</td>
<td>-0.937</td>
</tr>
<tr>
<td>Program*Training</td>
<td>4.375</td>
<td>2.187</td>
</tr>
<tr>
<td>Program*Criteria</td>
<td>1.875</td>
<td>0.937</td>
</tr>
<tr>
<td>Program*Num Peop</td>
<td>-1.875</td>
<td>-0.937</td>
</tr>
<tr>
<td>Experien*Training</td>
<td>-5.625</td>
<td>-2.812</td>
</tr>
<tr>
<td>Experien*Criteria</td>
<td>4.375</td>
<td>2.188</td>
</tr>
<tr>
<td>Experien*Num Peop</td>
<td>-4.375</td>
<td>-2.187</td>
</tr>
<tr>
<td>Training*Criteria</td>
<td>-1.875</td>
<td>-0.938</td>
</tr>
<tr>
<td>Training*Num Peop</td>
<td>1.875</td>
<td>0.937</td>
</tr>
<tr>
<td>Criteria*Num Peop</td>
<td>1.875</td>
<td>0.937</td>
</tr>
</tbody>
</table>

Analysis of Variance for % (coded units)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Seq SS</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Effects</td>
<td>5</td>
<td>2932.8</td>
<td>2932.8</td>
<td>586.56</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>2-Way Interactions</td>
<td>10</td>
<td>440.6</td>
<td>440.6</td>
<td>44.06</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Residual Error</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>3373.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Shows same thing
Training, Criteria, Experience
as the influential factors
Main Effects Plot (data means) for % Match

Same data but not we know whether to set high or low

Whoa, what do we have here?

Process pretty robust to program, number of reviewers
Interaction Plot (data means) for % Match

- Program
- Experience
- Training
- Criteria
- Num People
Experience was no brainier - more is better

Training was no brainier - have to have it!
  Phase dependent training, not just peer review training

Number of people was eye opening - didn’t make too much difference
  Since more people cost more money, keep it at 2

Criteria was a shock
  Need to do follow-up to see why this was counter-intuitive

Follow up revealed criteria limited the scope of the review, reviewed only what was there, did not use as intended
Other Measures Were Monitored, But Not Part of DOE
Thought Process Map - big benefit when reviewing project with others. Helps avoid heading down a dead end path before you even start. Makes you ask questions and identify barriers well before you actually get there. Much more critical than first thought. Is a pain to keep up to date

Process Map - get a feel for how the process is actually operating, not how it is designed to operate. Important to get a varying cross section of disciplines to get all point of views. Key to almost everything, need to know what x’s are involved in achieving Y.

FMEA - Identified possible failures, severity, occurrence and delectability, and prioritized the actions that should be taken. Important to get cross sectional input and consensus. Influential factors were hard to miss.

S/W Worksheet (Product Scorecard) - Collects and categorizes defect data across software life cycle. Facilitates to baseline data and track improvement. Unmodified SW Worksheet doesn’t tell you everything you need to show improvement.

Pareto Chart - Bar chart ordered from largest to smallest. Helped immediately determine which items to focus improvements upon for the largest payback and where to ask initial questions.

Control Charts - Distinguishes special and common variation in the process. Helps to develop appropriate action for the type of variation. Showed how process would perform if nothing done to improve it.

DOE - Determined what factors were influential and were to set those factors. Showed that the obvious accepted conclusions are not always an improvement.

Bottom line -
Until the DOE all the tools were useful and we could see their value. They confirmed what was already known. We were able to move forward confident that we were working the right issues and had the metrics to back up my assumptions. After the DOE a valuable lesson was learned. We do not know everything. If we hadn’t used the data to run the DOE, we would have made things worse!!
Results
### Number of defects identified by phase introduced/phase detected (from Product Scorecard)

<table>
<thead>
<tr>
<th>Phase</th>
<th>Planning</th>
<th>Customer</th>
<th>Rqmts. Analysis</th>
<th>Design</th>
<th>Implementation</th>
<th>Test</th>
<th>Formal Test</th>
<th>Customer Before</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>235</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>240</td>
</tr>
<tr>
<td>Customer</td>
<td>0</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Rqmts. Analysis</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>14</td>
<td>29</td>
<td>26</td>
<td>71</td>
<td>1</td>
<td>202</td>
</tr>
<tr>
<td>Design</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>323</td>
<td>82</td>
<td>29</td>
<td>38</td>
<td>2</td>
<td>475</td>
</tr>
<tr>
<td>Implementation</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>220</td>
<td>43</td>
<td>44</td>
<td>10</td>
<td>323</td>
</tr>
<tr>
<td>Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>249</td>
<td>30</td>
<td>0</td>
<td>282</td>
</tr>
<tr>
<td>Formal Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>597</td>
<td>0</td>
<td>610</td>
</tr>
<tr>
<td>Customer Before</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL</td>
<td>29</td>
<td>12</td>
<td>65</td>
<td>344</td>
<td>334</td>
<td>360</td>
<td>786</td>
<td>17</td>
<td>1947</td>
</tr>
</tbody>
</table>

**PHASE DETECTED**

<table>
<thead>
<tr>
<th>Phase</th>
<th>Planning</th>
<th>Customer</th>
<th>Rqmts. Analysis</th>
<th>Design</th>
<th>Implementation</th>
<th>Test</th>
<th>Formal Test</th>
<th>Customer Before</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>235</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>240</td>
</tr>
<tr>
<td>Customer</td>
<td>0</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Rqmts. Analysis</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>14</td>
<td>29</td>
<td>26</td>
<td>71</td>
<td>1</td>
<td>202</td>
</tr>
<tr>
<td>Design</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>323</td>
<td>82</td>
<td>29</td>
<td>38</td>
<td>2</td>
<td>475</td>
</tr>
<tr>
<td>Implementation</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>220</td>
<td>43</td>
<td>44</td>
<td>10</td>
<td>323</td>
</tr>
<tr>
<td>Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>249</td>
<td>30</td>
<td>0</td>
<td>282</td>
</tr>
<tr>
<td>Formal Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>597</td>
<td>0</td>
<td>610</td>
</tr>
<tr>
<td>Customer Before</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL</td>
<td>235</td>
<td>25</td>
<td>186</td>
<td>200</td>
<td>466</td>
<td>46</td>
<td>3</td>
<td>5</td>
<td>1166</td>
</tr>
</tbody>
</table>
### Percentage of defects identified by phase introduced/phase detected (from Product Scorecard)

<table>
<thead>
<tr>
<th>PHASE</th>
<th>Planning</th>
<th>Customer</th>
<th>Rqmts. Analysis</th>
<th>Design</th>
<th>Implementation</th>
<th>Test</th>
<th>Formal Test</th>
<th>Customer Before</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td>Customer</td>
<td>0%</td>
<td>57%</td>
<td>10%</td>
<td>0%</td>
<td>10%</td>
<td>0%</td>
<td>24%</td>
<td>0%</td>
<td>1%</td>
</tr>
<tr>
<td>Rqmts. Analysis</td>
<td>0%</td>
<td>0%</td>
<td>30%</td>
<td>7%</td>
<td>14%</td>
<td>13%</td>
<td>35%</td>
<td>0%</td>
<td>10%</td>
</tr>
<tr>
<td>Design</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>68%</td>
<td>17%</td>
<td>6%</td>
<td>8%</td>
<td>0%</td>
<td>24%</td>
</tr>
<tr>
<td>Implementation</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
<td>68%</td>
<td>13%</td>
<td>14%</td>
<td>3%</td>
<td>17%</td>
</tr>
<tr>
<td>Test</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
<td>88%</td>
<td>11%</td>
<td>0%</td>
<td>14%</td>
</tr>
<tr>
<td>Formal Test</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>2%</td>
<td>98%</td>
<td>0%</td>
<td>31%</td>
</tr>
<tr>
<td>Customer Before</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>20%</td>
<td>80%</td>
<td>0%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1%</td>
<td>1%</td>
<td>3%</td>
<td>18%</td>
<td>17%</td>
<td>18%</td>
<td>40%</td>
<td>1%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**AFTER IMPROVEMENTS**
### PHASE DETECTED

<table>
<thead>
<tr>
<th>PHASE</th>
<th>Planning</th>
<th>Customer</th>
<th>Rqmts. Analysis</th>
<th>Design</th>
<th>Implementation</th>
<th>Test</th>
<th>Formal Test</th>
<th>Customer Before</th>
<th>TOTAL</th>
<th>Leaked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>16.45</td>
<td>2</td>
<td>0</td>
<td>0.51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18.96</td>
<td>2.51</td>
</tr>
<tr>
<td>N</td>
<td>Customer</td>
<td>0</td>
<td>3.3</td>
<td>15.4</td>
<td>7</td>
<td>8.12</td>
<td>0</td>
<td>0</td>
<td>33.82</td>
<td>30.52</td>
</tr>
<tr>
<td>T</td>
<td>Rqmts. Analysis</td>
<td>0</td>
<td>1</td>
<td>19.32</td>
<td>11.18</td>
<td>7.91</td>
<td>3.2</td>
<td>0</td>
<td>42.61</td>
<td>23.29</td>
</tr>
<tr>
<td>R</td>
<td>Design</td>
<td>0</td>
<td>0</td>
<td>0.39</td>
<td>22.49</td>
<td>11.1</td>
<td>1.6</td>
<td>0</td>
<td>35.58</td>
<td>13.09</td>
</tr>
<tr>
<td>O</td>
<td>Implementation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>239.4</td>
<td>42</td>
<td>0</td>
<td>283.93</td>
<td>44.53</td>
</tr>
<tr>
<td>D</td>
<td>Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.16</td>
<td>0.06</td>
<td>1.76</td>
<td>0.15</td>
<td>0</td>
<td>2.13</td>
</tr>
<tr>
<td>U</td>
<td>Formal Test</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.58</td>
<td>1.08</td>
<td>0.58</td>
</tr>
<tr>
<td>C</td>
<td>Customer Before</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10.2</td>
<td>10.2</td>
<td>10.2</td>
</tr>
<tr>
<td>D</td>
<td>TOTAL</td>
<td>16.45</td>
<td>6.3</td>
<td>35.11</td>
<td>41.54</td>
<td>266.59</td>
<td>48.56</td>
<td>0.65</td>
<td>428.31</td>
<td>125.09</td>
</tr>
</tbody>
</table>

**Cost of Rework due to Defects in Days (from Product Scorecard)**

**AFTER IMPROVEMENTS**
95% Confident That Two of The Three Phase Will be World-Class. The Third Had Drastic Improvement

What happened here? Remember we are measuring % not#. We went from 282 to 27, a huge improvement in rework $ and effort, but % wise it was a decline

AFTER IMPROVEMENTS
AFTER IMPROVEMENTS
Bottom Line Savings vs Goal - Show Me the Money

$120,000 in six months on just 3 software projects

(Notice the increased up front spending)
Three Phase Where Project Concentrated (Three Phases With > 92% of Rework)

Recalculating the control limits to see if there is significance....
Control Limits for Requirements Analysis Phase has no Overlap Whatsoever

Design Phase is More Interesting

Control Limits for Implementation Phase Have Significantly Changed With Little Overlap

Range is much larger which widens the control limits.

However, looking at the scorecard,
- only 1% of defects made it to test, whereas before it was 14%.
- Rework due to leakage was a mere 13 days compared to 156 days!
Conclusion:

• Make decisions based on data.
  • Experience is only one input parameter and can steer you wrong

• The Six Sigma tools can and do apply to software

• A few simple process changes resulted had BIG impact to bottom line!

• The three phases that were concentrated upon improved greatly
  • Two of the three are above the threshold

• Quick look at charts show that the process is now in control

• Just measuring the percentage of defect leakage is not the whole story
  • Test injected defects were significantly reduced which resulted in a major cost savings
That’s All Folks

Tom Lienhard
Thomas_G_Lienhard@Raytheon.com
(520) 794-2989