On Human Analyst Performance in Assisted Requirements Tracing: Statistical Analysis

Alex Dekhtyar
Olga Dekhtyar
Jeff Holden
David Cuddeback

Jane Huffman Hayes
Wei-Keat Kong

Funded in part by NSF under grant CCF-0811140 and Lockheed Martin
Outline

- Assisted Requirements Tracing
- Human Analyst Performance
- Statistical Analysis
Requirements Traceability

«Requirements traceability is the ability to follow the life of a requirement in both backward and forward directions, i.e. from its origins, through its development and specification, to its subsequent deployment and use, and through periods of ongoing refinement and iteration in any of these phases» [Gotel’97]
Requirements traceability is the ability to follow the life of a requirement in both backward and forward direction, i.e. from its origins, through its development and specification, to its subsequent deployment and use, and through periods of ongoing refinement and iteration in any of these phases» [Gotel’97]
Desired Tracing Properties

- Fast
- Certifiable
- Accurate
Who performs tracing
Manual Tracing Process
Fully Automated Tracing
Assisted (Semi-Automated) Tracing
Best of Both Worlds!
What makes it accurate?
Candidate Traceability Matrix
Candidate Traceability Matrix

Test cases

Requirements

True Links

False Positives
Candidate Traceability Matrix

- Test cases
- Requirements
- True Links
- Missing Links
- False Positives
What makes it accurate?

True Links: can analysts validate them?
False Positives: can analysts reject them?
Missing Links: can analysts find them?
Big Question!

Can humans make correct decisions when performing assisted tracing???
What we knew before

• Pilot study (Hayes, Dekhtyar, MSR 2005)
• Semi-automated tracing
• Control initial TM accuracy
• Measure final TM accuracy
• Four data points
How we visualize

Recall (percent of true links found)

Precision

(Percent retrieved links that are true)

(1,1): True Trace

(1,1): True Trace

(Percent of true links found)
How we visualize

Recall (percent of true links found)

Precision

(Percent retrieved links that are true)

(1,1): True Trace

Final (submitted) Trace

Starting Trace

Recall (percent of true links found)
How we visualize

Recall (percent of true links found)

Precision (percent retrieved links that are true)

(1,1): True Trace

Final (submitted) Trace

Starting Trace

(0,0) (percent of true links found)
Desired behavior

(Percent retrieved links that are true)

Precision

Recall

(Percent of true links found)

(1,1): True Trace
Pilot Results
Cohorts

1. RETRO [Cuddeback et al., RE’2010] {33}

2. RETRO.net {13}

3. Manual – Hardcopy artifacts {38}
Tools - RETRO
2.1.10

Has the ability to format multiple files at the same time (i.e. all source code in a package.) Also have the ability to format just one file at a time.

TC-17

Purpose:

* User can add/modify Style Preferences through UI.

Procedure:

* Download and install "changepage.jar" from the LUNA Wiki Page.
* Create a new file called "JDialogPageConvention.xml" with the text from case #18.
* Create a new JDialog Project called "Test Case MARK".
* Create a new Class called "TestClass".
* Edit "TestClass" (see case #18).
* Go to Tools > Preferences...
* Go to the [Extensions] tab
* Click [Edit...]
* Navigate to the General section.
* Type "New Convention" in the "Name:" field.
* Type "New (Test) Convention" in the "Description:" field.
Pre-Study Survey

- Collect participants' experience levels
  - Demographic info (major, class, etc.)
  - SE courses
  - Industry experience
  - Tracing experience
  - Tracing confidence
Experiment Setup

• Assign user IDs by experience, quadrant
  – Specified precision & recall values

• Add/remove links from baseline
  – Baseline TM
    • 100% recall, 5% precision
Post-Study Survey

• Collect participant reactions
  – Reaction to the tool & task
  – Features used during task
  – Confidence during task
Dataset - ChangeStyle

• BlueJ Java code formatter
  – CSC 308/309 project
  – 32 requirements
  – 17 test cases
  – 24 true links

• True TM vetted by research team
  – Consensus of 6 people
Data Collected

• 3 types
  – Baseline independent variables
  – Observed independent variables
  – Dependent variables
Baseline Independent

- Candidate TM accuracy
  - Initial Precision
  - Initial Recall
  - Initial f2
  - Quadrant

- Candidate TM size
  - Number correct links
  - Number incorrect links
  - Total number of links
Observed independent Variables

- Data collected from Pre- and Post-experiment surveys
- 11 variables
  1. Method Used
  2. Location
  3. SE experience
  4. Tracing Experience
  5. Time to complete task
  1. Grade level
  2. Confidence in tracing
  3. Opinion on Tool vs. Manual
  4. Effort on omitted links
  5. Effort on offered links
  6. Preparedness
Dependent Variables

- Measure final TM accuracy
  - Final Precision
  - Final Recall
  - Final f2
  - Delta Precision
  - Delta Recall
  - Delta f2
Research Questions

• Is the effect of the baseline variables statistically significant on the dependent variables?

• Do any of the observed independent variables affect the dependent variables in a statistically significant way?

• What variable (or group of variables) has the largest statistically significant impact on the dependent variables?
What we observed

Starting TM Accuracy
What we observed

Submitted TM Accuracy

Precision vs. Recall plot showing the performance distribution for different models.
What possible insight?
Nobody Recovered True Traceability Relation
Good TMs in...

Degraded TMs out?
Bad TMs in...
Large TMs in...

Removed false positives...
...together with true links
Small TMs in...

Found missing links ...
... but added false positives
Statistical Analysis

• Report following:
  – $R^2$: How well the regression line approximates real data
    • [0 – 1] - higher is better predictor
  – F-value: How likely future data will be at least as random as observed data
  – Sig. (pval): Probability of obtaining results at least as extreme as observed
    • We use $pval < 0.05$ (95% or greater)
**Statistical Analysis**

**Initial Precision & Recall on Dependent Response Variable**

<table>
<thead>
<tr>
<th>Response Variable</th>
<th>( R^2_{\text{adj}} )</th>
<th>F-value</th>
<th>Sig. (pval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Precision</td>
<td>0.120</td>
<td>6.659</td>
<td><strong>0.002</strong></td>
</tr>
<tr>
<td>Final Recall</td>
<td>-0.004</td>
<td>0.842</td>
<td>0.434</td>
</tr>
<tr>
<td>Final ( f^2 )</td>
<td>0.000</td>
<td>1.012</td>
<td>0.368</td>
</tr>
<tr>
<td>Delta Precision</td>
<td>0.454</td>
<td>35.548</td>
<td><strong>0.0001</strong></td>
</tr>
<tr>
<td>Delta Recall</td>
<td>0.444</td>
<td>34.115</td>
<td><strong>0.0001</strong></td>
</tr>
<tr>
<td>Delta ( f^2 )</td>
<td>0.288</td>
<td>17.761</td>
<td><strong>0.0001</strong></td>
</tr>
</tbody>
</table>

Initial accuracy serves as a good predictor for final precision and all three deltas.
### Statistical Analysis

#### $f_2$ on Dependent Response Variable

<table>
<thead>
<tr>
<th>Response Variable</th>
<th>$R^2_{adj}$</th>
<th>F-value</th>
<th>Sig. (pval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Precision</td>
<td>0.056</td>
<td>5.931</td>
<td>0.017</td>
</tr>
<tr>
<td>Final Recall</td>
<td>0.037</td>
<td>3.117</td>
<td>0.081</td>
</tr>
<tr>
<td>Final $f_2$</td>
<td>0.053</td>
<td>4.604</td>
<td>0.035</td>
</tr>
<tr>
<td>Delta Precision</td>
<td>0.036</td>
<td>3.020</td>
<td>0.086</td>
</tr>
<tr>
<td>Delta Recall</td>
<td>0.312</td>
<td>37.227</td>
<td>0.0001</td>
</tr>
<tr>
<td>Delta $f_2$</td>
<td>0.238</td>
<td>25.672</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Initial $f_2$ acting as a one-dimensional surrogate can be a good predictor for final precision, final recall, delta recall, and delta $f_2$. 
## Statistical Analysis

### Initial Quadrant on Dependent Response Variable

<table>
<thead>
<tr>
<th>Response Variable</th>
<th>$R^2_{adj}$</th>
<th>F-value</th>
<th>Sig. (pval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Precision</td>
<td>0.138</td>
<td>5.434</td>
<td><strong>0.002</strong></td>
</tr>
<tr>
<td>Final Recall</td>
<td>0.004</td>
<td>1.113</td>
<td>0.349</td>
</tr>
<tr>
<td>Final $f_2$</td>
<td>0.038</td>
<td>2.083</td>
<td>0.109</td>
</tr>
<tr>
<td>Delta Precision</td>
<td>0.402</td>
<td>19.586</td>
<td><strong>0.0001</strong></td>
</tr>
<tr>
<td>Delta Recall</td>
<td>0.341</td>
<td>15.344</td>
<td><strong>0.0001</strong></td>
</tr>
<tr>
<td>Delta $f_2$</td>
<td>0.253</td>
<td>10.356</td>
<td><strong>0.0001</strong></td>
</tr>
</tbody>
</table>

Initial quadrant acting as a coarse one-dimensional surrogate can be a good predictor for final precision as well as all three deltas. Confirms informal findings that behavior is tied to quadrant.
Statistical Analysis - Observed Independent Variables

Only one had statistical significance: Effort Validating Candidate Links
Effect of Validating Link

- Analysts who applied most effort:
  - Overthinking?
  - Reduced Recall
What affects final accuracy and change in accuracy?

Yes

Initial Accuracy (Recall, Precision)

Effort to confirm/reject presented candidate links

No

Location  Opinion  Time to complete
Process  Grade level  Confidence
SE experience  Tracing experience  Preparedness
Big Picture

Analysts are fallible

Accurate automated methods not of much help? (analyst yields best TM from lower quality initial TM – must account for this)

... but tracing MUST be done

Need follow-on with data from real project(s) to study this behavior
Questions?
## Trace validation experiment

<table>
<thead>
<tr>
<th>Three Experiments</th>
<th>Two Sites</th>
<th>Small Simple Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>UI 1 (RETRO)</td>
<td>Cal Poly</td>
<td>17 requirements</td>
</tr>
<tr>
<td>UI 2 (RETRO.NET)</td>
<td>University of Kentucky</td>
<td>32 test cases</td>
</tr>
<tr>
<td>Manual Trace validation</td>
<td></td>
<td>24 true links</td>
</tr>
<tr>
<td>{33}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{13}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>{38}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Trace validation experiment

## Three Experiments

<table>
<thead>
<tr>
<th>Experiment</th>
<th>UI 1 (RETRO)</th>
<th>UI 2 (RETRO.NET)</th>
<th>Manual Trace validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>{33}</td>
<td>{13}</td>
<td>{38}</td>
</tr>
</tbody>
</table>

## Two Sites

- Cal Poly
- University of Kentucky

## Small Simple Dataset

- 17 requirements
- 32 test cases
- 24 true links

## Controlled For

- Initial TM precision
- Initial TM recall

## Task

- Validate given candidate TM

---

52
# Trace validation experiment

<table>
<thead>
<tr>
<th>Three Experiments</th>
<th>Two Sites</th>
<th>Small Simple Dataset</th>
<th>Task</th>
<th>Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>UI 1 (RETRO)</td>
<td>Cal Poly</td>
<td>17 requirements</td>
<td>Validate given candidate TM</td>
<td>Final TM precision</td>
</tr>
<tr>
<td>UI 2 (RETRO.NET)</td>
<td>University of Kentucky</td>
<td>32 test cases</td>
<td></td>
<td>Final TM recall</td>
</tr>
<tr>
<td>Manual Trace validation</td>
<td></td>
<td>24 true links</td>
<td></td>
<td>53</td>
</tr>
</tbody>
</table>

**Controlled For**
- Initial TM precision
- Initial TM recall

**Collected**
- 11 parameters
  - (experience, opinions, effort, confidence...)

**Parameters**
- Initial TM precision
- Initial TM recall