Programmers are Users Too: Human Centered Methods for Improving Tools for Programming

Brad A. Myers
Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
http://www.cs.cmu.edu/~bam
bam@cs.cmu.edu
Natural Programming Project

• Researching better tools for programming since 1978
• Natural Programming project started in 1995
• Make programming easier and more correct by making it more natural
  – Closer to the way that people think about algorithms and solving their tasks
• Methodology - human-centered approach
  – Perform studies to inform design
    • Provide new knowledge about what people do and think, & barriers
    • Guide the designs from the data
  – Design of programming languages and environments
  – Iteratively evaluate and improve the tools

Target novice, expert and end-user programmers
End User Programming

- People whose primary job is not programming
- [Scaffidi, Shaw and Myers 2005]
  - 90 million computer users at work in US
  - 55 million will use spreadsheets or databases at work (and therefore may potentially program)
  - 13 million will describe themselves as programmers
  - 3 million professional programmers
- All of these people use APIs!
Goal: Gentle Slope
Systems

Program Complexity and Sophistication

Goal

Low Threshold

High Ceiling

Difficulty of Use

Web Development

Java

Visual Basic

Flash

Server-side

C Programming

C or C# Programming

Swing

JavaScript

ActionScript

CSS & HTML

Basic

Email

Editor

4 © 2017 – Brad A. Myers
Human Centered Approaches?

• Concerned with everything the user encounters
  – Functionality & Usefulness
  – Content
  – Labels
  – Presentation
  – Layout
  – Navigation
  – Speed of response
  – Emotional Impact
  – Context (social environment in which use happens)
  – Documentation & Help
• Measures:
  – Learnability, Productivity, Errors, …
What Can Be Addressed?

- **Everything** the developer encounters
- **Tools** – IDEs & their user interfaces
- **Languages themselves**
  - Not necessarily just “taste”, “intuition”
  - Error-proneness
- **APIs**
  - “Interface” between developer and functionality
  - “Languages” by themselves are almost irrelevant these days
- **Documentation for all of the above**
- **Processes & context of development**
  - Consider the whole “system” together
  - New as well as legacy systems
“Human Centered Approaches” – More Than Lab User Studies

• Design & aesthetics matter & will affect:
  – User’s performance
  – Errors
  – Adoption of your tool

• Many different methods for answering many different questions
  – Before design time
  – During design & implementation
  – After implementation
Many HCI Methods

- Contextual Inquiry
- Contextual Analysis
- Paper prototypes
- Think-aloud protocols
- Heuristic Evaluation
- Affinity diagrams
- Personas
- Wizard of Oz
- Task analysis
- A/B testing
- Cognitive Walkthrough
- Cognitive Dimensions
- KLM and GOMS (CogTool)
- Video prototyping
- Body storming
- Expert interviews
- Questionnaires
- Surveys
- Interaction Relabeling
- Log analysis
- Storyboards
- Focus groups
- Card sorting
- Diary studies
- Improvisation
- Use cases
- Scenarios
- “Speed Dating”
- …

© 2017 – Brad A. Myers
Dangers of *Not* Applying Human Centered Approaches

- Tools may prove to be not useful
  - Useful = solves an important problem
    - Happens frequently
    - Difficult to solve otherwise
    - Developers believe academic tools solve unimportant problems

- Tools may not actually solve the problem
  - Example: a study suggested that Tarantula tool identifying potentially faulty statements for debugging was not helpful
    - Changed the task, but telling if the identified statement was actually faulty not easier than finding the bug
Dangers of *Not* Applying Human Centered Approaches

- Tools may show no measurable impact
  - Desired advantage overwhelmed by problems with other parts
  - Example: Emerson Murphy-Hill found that refactoring tools are under-utilized and programmers do not configure them due to usability issues

Human Centered Approaches are Not Too Difficult for You

- Getting *some* user data better than none
- Observing real usage reveals many opportunities
  - Insights about new issues to address, not necessarily what originally planned
    - Thomas LaToza’s Reachability Questions from Architecture study
    - Jeff Stylos’s method placement result from study of class size: from 2.4 to 11.2 times faster
      
      ```
      server.send ( message ) vs.
      mail.send ( server )
      ```
    
- Collaborating with Graphic Designers for even a short time can provide significant improvements in aesthetics
Key Decision: What is Your Question?

- What do you need to find out or show?
  - What claim do you want to make?
- Showing that a tool is usable is different from whether it is useful.
- Exploring what people are doing, is different from determining how often an observed behavior happens.
  - Drives what type of method to use, and tasks to be done with it.
Product Lifecycle

Source: http://www.accordtech.co.in/Product%20Development%20Lifecycle.htm
Product Lifecycle

Field Studies
- Logs & error reports

Evaluative Studies
- Expert analyses
- Usability Evaluation
- Formal A/B Lab Testing

Exploratory Studies
- Contextual Inquiries
- Interviews
- Surveys
- Lab Studies
- Corpus data mining

Design Practices
- "Natural programming"
- Graphic & Interaction Design
- Prototyping
Exploratory Studies

- Identify what is really happening
- Discover important problems
- Quantify need
Contextual Inquiry


- A kind of “ethnographic” or “participatory design” method
- Watch developers while they are performing their real tasks
- Objective, concrete data about real activities
- May be followed by a survey, to establish generality of the issues
Why Contextual Inquiry?

• Usually reveals many barriers and problems in current practice
• Helps develop *insights*
  – Be open to inspiration
• Not for confirming what you already know
• Qualitative data (not quantitative)
  – CIs are not for gathering statistics, analytics
    • In contrast to surveys & lab studies
• But need to be able to observe real tasks
Example of Contextual Inquiry & Surveys

“Developers Ask Reachability Questions”
- “Search across feasible paths through a program for target statements matching search criteria”
  - Watched 17 developers investigating unfamiliar code
  - Also surveyed 460 developers
  - Over 100 other hard-to-answer questions
Many hard-to-answer questions about code (PLATEAU'2010)

Rationale (42)
Why was it done this way? (14) [15][7]
Why wasn’t it done this other way? (15)
Was this intentional, accidental, or a hack? (9)[13]
How did this ever work? (4)

Debugging (26)
How did this runtime state occur? (12) [15]
What runtime state changed when this executed? (2)
Where was this variable last changed? (1)
How did I debug this bug in this environment? (3)
In what circumstances does this bug occur? (3) [15]
Which team’s component caused this bug? (1)

Intent and Implementation (32)
What is the intent of this code? (12) [15]
What does this do (6) in this case (10)? (16) [24]
How do I build this (8), given this constraint (2)? (10)
Which function or object should I pick? (2)
What’s the best design for implementing this? (7)

Control flow (19)
In what situations or user scenarios is this called? (3) [15][24]
What parameter values does each situation pass to this method? (1)
What parameter values could lead to this case? (1)
What are the actual methods called by dynamic dispatch here? (6)
How do calls flow across process boundaries? (1)
How many recursive calls happen during this operation? (1)
What is the correct order for calling these methods or initializing these objects? (2)

Contracts (17)
What assumptions about preconditions does this code make? (5)
What assumptions about post?conditions can be made? (5)
What exceptions or errors can this method generate? (2)
What are the constraints on or normal values of this variable? (2)
What is the correct order for calling these methods or initializing these objects? (2)
Should I refactor this? (1)
Are the benefits of this refactoring worth the time investment? (3)

Performance (16)
What is the performance of this code (5) on a large, real dataset? (3)
Which part of this code takes the most time? (4)
Can this method have high stack consumption from recursion? (1)
How big is this in memory? (2)
How many of these objects get created? (1)

Type relationships (15)
What are the composition, ownership, or usage relationships of this type? (5) [24]
What is this type’s type hierarchy? (4) [24]
What implements this interface? (4) [24]
Where is this method overridden? (2)

Data flow (14)
What is the original source of this data? (2) [15]
What code directly or indirectly uses this data? (5)
What is the data referenced by this variable modified? (2)
Where can this global variable be changed? (1)
What is this data structure used for (1) for this purpose? (1) [24]
What parts of this data structure are modified by this code? (1) [24]
What resources is this code using? (1)

Location (13)
Where is this functionality implemented? (5) [24]
Is this functionality already implemented? (5) [15]
Where is this defined? (3)

Building and branching (11)
Should I branch code against the main branch? (1)
How can I move this code to this branch? (1)
What do I need to include to build this? (3)
What includes are unnecessary? (2)
How do I build this without doing a full build? (1)
Why did the build break? (2)[59]
Which preprocessor definitions were active when this was built? (1)

Architecture (11)
How does this code interact with libraries? (4)
What is the architecture of the code base? (3)
How is this functionality organized into layers? (1)
Is my API understandable and flexible? (3)

History (23)
When, how, by whom, and why was this code changed or inserted? (15)
What else changed when this code was changed or inserted? (2)
How has this changed over time? (4)[7]
Has this code always been this way? (2)
What recent changes have been made? (1)[15][7]
Have changes in another branch been integrated into this branch? (1)

Concurrency (9)
What threads reach this code (4) or data structure? (2)[6]
Is this class or method thread-safe? (2)
What members of this class does this lock protect? (1)

Dependencies (5)
What depends on this code or design decision? (4)[7]
What does this depend on? (1)

Method properties (2)
How big is this code? (1)
How overloaded are the parameters to this function? (1)

Teammates (16)
Who is the owner or expert for this code? (3)[7]
How do I convince my teammates to do this the “right way”? (12)
Did my teammates do this? (1)

Policies (15)
What is the policy for doing this? (10) [24]
Is this the correct policy for doing this? (2) [15]
How is the allocation lifetime of this object maintained? (3)

Implications (21)
What are the implications of this change for (5) API clients (5), security (3), concurrency (3), performance (2), platforms (1), tests (1), or obfuscation (1)? [21][15][24]
Many opportunities for better tools

- Of all the reported questions
  - 34% addressed by commercial tools
  - 25% addressed by research tools
  - 41% unaddressed by any tools
Example of Interviews: Immutability

- Experts recommend making classes *immutable* so instances cannot change accidentally
  - Thread safe, more secure, no unexpected state changes, etc.
- Usability studies suggest programmers prefer classes that *can* change
- Various relevant language features
  - C++ `const`, Java `final`, Obj-C immutable collections, .NET `Freezable`, etc.
- Semi-structured interviews with a convenience sample of 8 software engineers
  - Agreed that mutability is a frequent source of bugs
  - But *none* of these features are what is needed
  - Preferred *transitive, class-based immutability*
    - Provide this in the **Glacier** tool *(to appear in ICSE’2017)*
    - **Great Languages Allow Class Immutability Enforced Readily**
Exploratory Lab Studies

- To understand what is happening
- More controlled than field studies
  - Can compare multiple people on same tasks
- Example: studying Eclipse for maintenance tasks
  - Detailed study of fixing bugs and adding features
  - Dataset used for 3 different award-winning papers: interruptions, navigation, code editing behaviors

<table>
<thead>
<tr>
<th>Interactive Bottleneck</th>
<th>Overall Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigating to fragment in <em>same</em> file (<em>via scrolling</em>)</td>
<td>~ 11 minutes</td>
</tr>
<tr>
<td>Navigating to fragment in <em>different</em> file (<em>via tabs and explorer</em>)</td>
<td>~ 7 minutes</td>
</tr>
<tr>
<td>Recovering working set after returning to a task</td>
<td>~ 1 minute</td>
</tr>
</tbody>
</table>

Total Costs ~19 minutes

=35%
Corpus Data Mining

- Studied 11 million Java try/catch blocks from GitHub using Boa tool
- 12% of catch blocks were completely empty.
- 25% of all exceptions caught are simply Exception
- Motivated a new tool to help programmers write better exception handling code

Figure 3: Exceptions caught by catch blocks on GitHub. Exceptions that occur more than 1% of the time are labeled. The rest, in purple, are thousands of exceptions that only rarely occur.
Design Methods

• Now know the problem, what is the solution?
• How do I design it so it is attractive and effective?
“Natural Programming”

• Technique developed by my group to elicit developer’s “natural” expressions
  – Mental models of tasks, vocabulary, etc.
• Blank paper tests
• Must prompt for the tasks in a way that doesn’t bias the answers
• Examples:
  – PacMan before and after
    • Mostly rule-based (if-then)
  – API designs
    • Architecture, words used, which methods are on which classes
Why Natural Programming?

- When want design to be easily learned by novices
- But biased by what they already know
  - Graphic designers will think PhotoShop is “natural”
  - Programmers will think Java is “natural”
Graphic Design

- Importance of graphic design and interaction design
- Software Engineers (and researchers) are not necessarily the best interaction designers
- Design can have a big impact even with same functionality
- Might involve designers for colors, icons, which controls, layout, …
- InterState system combines state transition diagrams and spreadsheets.
  - Carefully designed to be usable and attractive
  - Good graphic design
  - Animations on change

Prototyping

- Try out designs with developers before implementing them
  - **Paper**
    - "Low fidelity prototyping"
    - Often surprisingly effective
    - Experimenter plays the computer
    - Drawn on paper → drawn on computer
  - **Implemented Prototype ("Click through")**
    - Balsamiq, Axure, PowerPoint, Web tools (even for non-web UIs)
    - (no database)
  - **Real system**

- Need to test these with users!
- Better if sketchier for early design
  - Use paper or "sketchy" tools, not real widgets
  - People focus on wrong issues: colors, alignment, labels
  - Rather than overall structure and fundamental design
Example of Early Prototyping

- Thomas LaToza designing new visualization tool to try to help answer Reachability Questions
- Prototypes created with Omnigraffle and printed
- Revealed significant usability problems that were fixed before implementation
  - Graphical presentation
  - Controls
Another Example: Variolite

• How to support data scientists with exploratory programming?
• What kind of version control support would be useful?
  – Interviews and CI's showed that conventional approaches like Git are too heavy-weight
• Showed dozens of sketches to target users to get feedback on which seemed usable and useful
• Final design to appear at CHI ’2017
• Variations Augment Real Iterative Outcomes Letting Information Transcend Exploration
Evaluation Methods

• Does my tool work?
• Does it solve the developer’s problems?
• “If the user can’t use it, it doesn’t work!”
  – Susan Dray
Expert Analyses

• Usability experts evaluating designs to look for problems
  – Heuristic Analysis – [Nielsen] set of guidelines
  – Cognitive Dimensions – [Green] another set
  – Cognitive Walkthroughs – evaluate a task

• Can be inexpensive and quick
• However, experienced evaluators are better
  – 22% vs. 41% vs. 60% of errors found [Nielsen]

• Disadvantage: “just” opinions, open to arguments
Our Use of Expert Analyses

• Collaborating with SAP on their APIs and tools
• We studied SAP’s Enterprise Service-Oriented Architecture (eSOA) APIs & Documentation
  

• Naming problems:
  – Too long: MaterialSimpleByIDAndDescriptionQueryMessage_syncMaterialSimpleSelectionByIDAndDescriptionSelectionByMaterialDescription
  – Not understandable

FindCustomer.
Usability Evaluations

• Different from formal A/B “user testing”
  – Understand usability issues
  – Should be done early and often
    • Doesn’t have to be “finished” to let people try it

• “Think aloud” protocols
  – “Single most valuable usability engineering method”
    -- [Nielsen]
  – Users verbalize what they are thinking
    • Motivations, why doing things, what confused about
  – Don’t need many users
Why Usability Analysis

• Improve the user interface prior to:
  – Deployment
  – A/B testing (as a “pilot” test)

• Demonstrate that users can use the system
  – Show that novel features of the UI are understandable
Example of Using Usability Analysis

• Thomas LaToza’s REACHER tool for Reachability Questions went through multiple iterations
  – Revised based on paper prototype (discussed already)
  – Revised based on 1st evaluation of full system
    • E.g., replaced duplicates of calls to methods with pointers
    • Changed to preserve order of outgoing edges
    • Redesign of icons, interactions
Another Example of Usability Analysis

- **Sugilite: Smartphone Users Generating Intelligent Likeable Interfaces Through Examples**
  - Allow end-users to create automations on Smartphones
  - Initiate with speech commands
  - Record scripts by example
  - Generalizes from one or more examples

- 19 participants attempted 5 tasks
  - All completed at least 2 tasks successfully
  - 8 (42.1%) succeed in all 4 tasks
  - Overall, 65 out of 76 (85.5%) scripts worked
  - Feedback on what we need to improve
Formal A/B testing

- Formal A/B lab user tests are “gold standard” for academic papers – to show something is better
- But many issues in the study design
  - “Confounding” factors which were not controlled and are not relevant to study, but affect results
  - Tasks or instructions are mis-understood
  - Use prototypes & pilot studies to find these
- Statistical significance doesn’t mean real savings
- Be sure to collect qualitative data too
  - Strategies people are using
  - Why users did it that way
  - Especially when unexpected results
Example of A/B testing

- User testing of InterState compared to JavaScript

```
var ms_untiladvance;
window.setInterval(function()
    ms_until_advance = 5000 - $
    if (diff <= 0) {
        set_selected_index((sel
        reset_timer();
    }
```

![Diagram showing time taken (minutes) for Task 1 and Task 2 for JavaScript and InterState with p-values of p < .01** and p < .05 * indicating smaller is better.](image)

JavaScript

InterState

*p < .01**

*p < .05 *
Another Example of A/B testing

- **Gneiss: Gathering Novel End-user Internet Services using Spreadsheets**

- Novel spreadsheet interface for investigating hierarchical (e.g., JSON) data
  - Investigate using conventional spreadsheet formulas and drag-and-drop of columns

- Gneiss users significantly outperformed Excel users and programmers (p<.001)
A/B Testing of Programming Language Feature

- Glacier immutability extension to Java
- 20 experienced Java programmers
- Compared to using Java `final` as instructed by Josh Bloch

<table>
<thead>
<tr>
<th></th>
<th>final</th>
<th>Glacier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users who made errors enforcing immutability (after all tasks)</td>
<td>10/10</td>
<td>0/10</td>
</tr>
<tr>
<td>Completed FileRequest.execute() tasks with security vulnerabilities</td>
<td>4/8</td>
<td>0/8</td>
</tr>
<tr>
<td>Completed HashBucket.put() tasks with bugs</td>
<td>7/10</td>
<td>0/7</td>
</tr>
</tbody>
</table>
Field Studies of System in Use

- Find out what happens when the tool is really used
- Requires significant effort to make the tool sufficiently solid
Logging Actual Use

• Easier if instrument your tools
• Objective use data better than users’ recollections and opinions
• Many levels of data can be collected
  – Privacy issues
• Example: Fluorite logger for Eclipse
  – Fluorite: Full of Low-level User Operations Recorded In The Editor
  – Records all edits and events, including scrolling operations & source code,
  – Necessary to identify patterns of backtracking
Why Field Studies?

- Understand which features are used and how
  - Not necessarily *why*
  - Can sometimes follow up with questionnaires, interviews of actual users
  - Developers often are surprised at how system is used
- Demonstrate that people choose to use the system when optional
- Easy to instrument web systems, some on-line tools
Example of Field Analysis

- **Apatite**: Associative Perusing of APIs That Identifies Targets Easily

- Novel documentation tool that works *by association*
  - E.g., methods often used together

- Can start with verbs (actions) and find what classes implement them

- Couldn’t figure out a comparison tool or a lab study

- Deployed on the web

- Mostly used for fast lookup from partial names
Summary of Insights

• Field and lab studies can reveal the real questions
  – Answering these questions creates tools that are actually useful
• Researcher’s intuitions about what might be useful may be wrong
• Our experience highlights:
  – Developers often have specific questions in mind, which can be exploited in tools
  – Code views are central
  – Visualizations are often useful as navigation aides for code
  – Ability to search is key
    • Not just through code, but also through dynamic and static call-graphs, through time, etc.
More on This Topic

- **CHASE** and **USER** workshops at ICSE; **PLATEAU** at SPLASH/OOPSLA; **VL/HCC**


- Reading list for “**Human Aspects of Software Development (HASD)**” by Thomas LaToza and Brad Myers
  
Acronyms are fun!

And there are lots of Gemstones!!

- **Fluorite**: Full of Low-level User Operations Recorded In The Editor
- **Azurite**: Adding Zest to Undoing and Restoring Improves Textual Exploration
- **Euclase**: Eclipse Users’ Keystrokes Lessened by Attaching from Samples
- **Variolite**: Variations Augment Real Iterative Outcomes Letting Information Transcend Exploration
- **Apatite**: Associative Perusing of APIs That Identifies Targets Easily
- **Jadeite**: Java API Documentation with Extra Information Tacked-on for Emphasis
- **Jasper**: Java Aid with Sets of Pertinent Elements for Recall
- **Crystal**: Clarifications Regarding Your Software using a Toolkit, Architecture and Language
- **Mica**: Makes Interfaces Clear and Accessible
- **Aquamarine**: Allowing Quick Undoing of Any Marks And Repair Improving Novel Editing
- **Gneiss**: Gathering Novel End-user Internet Services using Spreadsheets
- **Sugilite**: Smartphone Users Generating Intelligent Likeable Interfaces Through Examples
- **GARNET**: Generating an Amalgam of Real-time, Novel Editors and Toolkits
- **Pebbles**: PDAs for Entry of Both Bytes and Locations from External Sources
- **Gneiss**: Gathering Novel End-user Internet Services using Spreadsheets
- **For more, see**: [www.cs.cmu.edu/~bam/acronyms.html](http://www.cs.cmu.edu/~bam/acronyms.html)
Thanks to:

- **Funding:**
  - NSF under IIS-1116724, IIS-0329090, CCF-0811610, IIS-0757511 (Creative-IT), ITR CCR-0324770 as part of the EUSES Consortium
  - SAP
  - Adobe
  - IBM
  - Microsoft Research RISE
  - Yahoo! InMind

- **>36 students:**
  - Htet Htet Aung
  - Jack Beaton
  - Ruben Carbonell
  - John R. Chang
  - Kerry S. Chang
  - Polo Chau
  - Luis J. Cota
  - Michael Coblenz
  - Dan Eisenberg
  - Brian Ellis
  - Andrew Faulring
  - Aristiwidya B. (Ika) Hardjanto
  - Erik Harpstead
  - Amber Horvath
  - Sae Young (Sophie) Jeong
  - Mary Beth Kery
  - Andy Ko
  - Thomas LaToza
  - Joonhwan Lee
  - Toby Li
  - Leah Miller
  - Steven Moore
  - Mathew Mooty
  - Gregory Mueller
  - Yoko Nakano
  - Stephen Oney
  - John Pane
  - Sunyoung Park
  - Michael Puskas
  - Chotirat (Ann) Ratanamahatana
  - Christopher Scaffidi
  - Jeff Stylos
  - David A. Weitzman
  - Yingyu (Clare) Xie
  - Zizhuang (Zizzy) Yang
  - YoungSeok Yoon
Programmers are Users Too: Human Centered Methods for Improving Tools for Programming

Brad A. Myers
Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
http://www.cs.cmu.edu/~bam
bam@cs.cmu.edu