Testing Spreadsheets

WYSIWYT Testing in the Spreadsheet Paradigm: An Empirical Evaluation

Presenter: Alex Aris

Overview

- Almost no work
  - SE tasks in creation and maintenance
- Real-world issues
  - Budgets, student grades, tax calculations
- Different from common PL
  - Declarative
  - Dependence-driven
  - Direct-manipulation working model

Do spreadsheets contain faults?

- 4 field audits
  - Errors in 20.6%
- 11 experiments (participants created…)
  - Errors in 60.8%
- 4 experiments (participants inspected…)
  - Missed 55.8% of errors

Motivation for WYSIWYT

- Why so many faults?
  - Overconfidence
  - Too much feedback and responsiveness
    - Interferes w/ problem solving

      GILMORE & SVENDSEN
- Feedback of testedness

Designed with … in mind

- Declarative evaluation model of spreadsheets formulas
- Incremental style of development
- Immediate visual feedback
- Various users

Points of Concern

- Is it efficient?
  - Coexist with the immediate redisplay after edit
  - Most algorithms O(1)
- Will faults be uncovered?
  - Hidden data-flow test adequacy criterion
- Will it decrease overconfidence?
  - Empirical studies

Features of Methodology

- Evaluation is driven by data dependencies
- Control flow only within cell formulas
- Flexibility
  - Scheduling algorithms
  - Optimization to computations
  - Incremental development
- Efficiency
  - Immediate visual response

Test Adequacy Criterion

- Can it be specification based?
  - Users not likely to write specs
- Code based testing adequacy
  - Output-influencing-all-du-criterion
    - "du adequacy"
    - Executable def-use's
    - Impossible/infeasible to determine by computation
    - No enforcement of evaluation order

Experiment Design

- More … than ad hoc?
  - Effective
    - DU adequacy
  - Efficient
    - Less redundancy
  - Less overconfident
- Training ?

Experimental Design

- People tested spreadsheets
- Experimental group
  - Includes WYSIWYT
- Control group
  - No WYSIWYT
- Recorded to transcript files
- Questionnaires
  - Subject background
  - Post-experiment
  - Use/understanding of WYSIWYT feedback

Experimental Environment

- Forms/3
  - Cells
    - Value defined by formula
- Grade book
- Visual clock with hands

Visual Feedback

- Cell Borders
  - Blue (fully tested)
  - Red (not tested)
  - Shades of purple
- Cell's check box
  - '✍' Tested
  - '?' Not fully tested
  - ' ' Further testing doesn't increase coverage
- %Tested indicator

Subjects

- Computer Science students
  - Experienced
    - 2 upper division undergraduate, 1 graduate
  - Potentially less room for improvement
  - No previous exposure

Groups

- Ad-hoc & WYSIWYT
- Random division
  - Subject to balancing grad & undergrad
- (37,41) ✍ (30,39)
  - Whose computer crashed
  - Who corrupted their data by mistake

Characteristics of Groups

- 0.2-0.25 lower GPA in WYSIWYT group
  - 1/3rd didn't report
  - Significant, but tenuous (slight)
  - Higher GPA assumed to lead in better performance
- Subjects w/ spreadsheet experience
  - 12/30 (ad-hoc) vs. 10/39
- Grad students
  - 10/30 (ad-hoc) vs. 8/39
- Professional experience
  - 11/30 (ad-hoc) vs. 20/39

Tutorial

- Quick reference handout
- 20-minute Forms/3 tutorial
  - Language features
    - Basic syntax of formulas
  - Environmental features
    - How to edit cells
- How to record testing decisions
  - Input cells, checking output cells
  - Incorrect cells ✍ Bug Recorder
- How to interpret the testing feedback

Training

- Total time: equal
- No info about du coverage

Tasks

- Testing Clock & Grades (both groups)
  - Different problem domains
    - Numeric vs. graphical
  - Clock difficult, grades easy to understand
  - Verifying
    - Difficult for Grades
    - Easy for Clock
- Familiar problems, limited time (15 min)
- Counterbalancing first & second spreadsheets

Results

- Effectiveness ⇙ du-adequacy
- Analysis of Variance (ANOVA)
  - Environment (WYSIWYT & ad-hoc)
  - Problem (Clock & grades)
- Significant difference
  - Effectiveness (du) & efficiency (wasted effort)
- No interaction effect

Results (continued)

- Speed as efficiency
  - Three 5-minute intervals
  - WYSIWYT subjects significantly faster after on the 3rd interval
- Overconfidence
  - Compare asked & calculated grades
    - Ad-hoc group was significantly more overconfident

Overconfidence

- Clock
  - Ad-hoc ⇙ 16/30 overconfident
  - WYSIWYT ⇙ 10/39 overconfident
- Grades
  - Ad-hoc ⇙ 20/30 overconfident
  - WYSIWYT ⇙ 14/39 overconfident

Redundancy

- Clock
  - Ad-hoc ⇙ 61.3% redundant
  - WYSIWYT ⇙ 15.4% redundant
- Grades
  - Ad-hoc ⇙ 44.0% redundant
  - WYSIWYT ⇙ 4.3% redundant

Helpfulness

- Feature: Very helpful, helpful, not helpful
  - Question marks 69% 31% 0%
  - Clicking to validate 64% 36% 0%
  - Colored cell borders 56% 44% 0%
  - Colored arrows 51% 41% 8%
  - Check marks 44% 49% 8%
  - 'Tested' indicator 36% 56% 8%
  - Blanks 23% 51% 26%

Understanding

- Opinions would be misleading if they didn't understand their meaning
- Had only 20 minutes to learn
- Asked 3 questions about meanings
  - Q1: 100%
  - Q2: 87%
  - Q3: 64%

Learning curve

- Even in the 1st problem WYSIWYT did better than Ad-hoc
- In the 2nd problem WYSIWYT benefited from experience
  - ~same number of test cases
  - Increased coverage (by 15%)
- Ad-hoc didn't significantly increase their coverage

Learning Effects

Problem 1	Tested	# Tests	Redundant
Ad-hoc	69.0%	13	51.3%
WYSIWYT	82.7%	20	11.1%

Problem 2	Tested	# Tests	Redundant
Ad-hoc	71.6%	22	56.3%
WYSIWYT	97.8%	18	7.7%

Threats to Validity

- Internal, addressed
  - Balanced 2 groups (year/class)
  - Counterbalancing problem type
  - Equalizing training time
  - Problems from familiar domains
- External, not addressed
  - CS students may not represent general population
  - Spreadsheets may not be representative enough

Threats to external validity (continued)

- WYSIWYT doesn't handle non-executable du associations
  - Avoided as much as possible
- No faults, formulas unchanged
  - Task would be interrupted
- Other measures for testing effectiveness
  - Number of faults detected
    - Also poses a threat to validity

Conclusion

- WYSIWYT subjects performed significantly better in terms of
  - Effectiveness
  - Efficiency
  - Being less overconfident
- Without formal training on the underlying testing theory

Discussion

Questions & Comments