Maximising the Information Gained From an Experimental Analysis of Code Inspection and Static Analysis for Concurrent Java Components

Margaret Wojcicki and Paul Strooper
School of Information Technology and Electrical Engineering
University of Queensland

13th September 2006

Background and Related Work

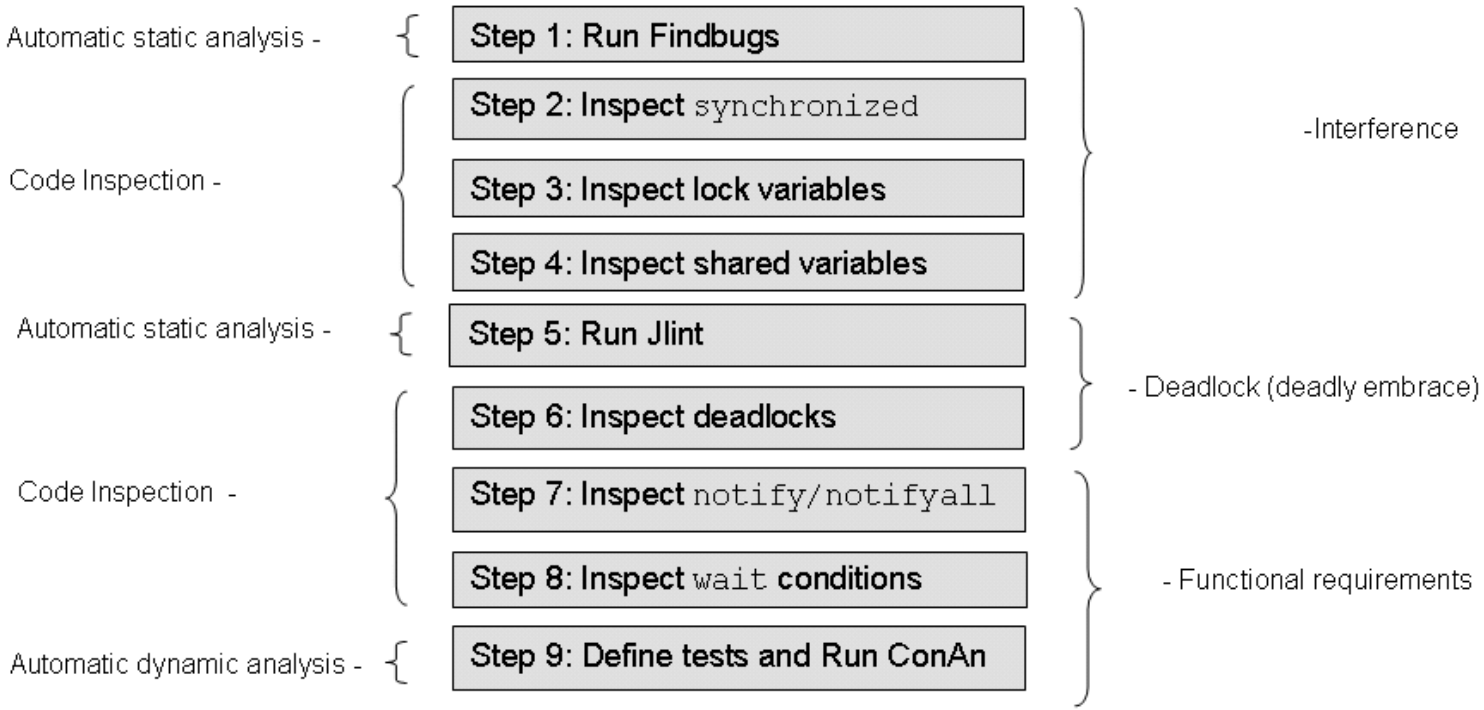
Verification and Validation

- Verification
  - does a computer program conform to its specification?
- Validation
  - does a computer program meet the requirements of the client?
- Verification and Validation (V&V):
  - dynamic
  - static
- Focus on V&V of components

Concurrent Components

- Concurrent components
  - two or more processes cooperate in performing a task
- Specific concurrent component defects
- Concurrent components are non-deterministic
- V&V technology is diverse

TestCon (Java)



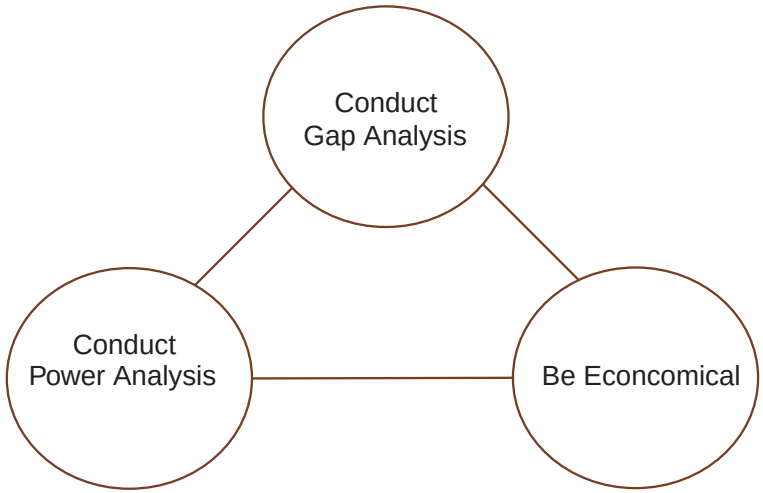
Related Studies

- Comparison (Basili & Selby, Kamsties & Lott) and combination (Selby and Wood et al.)
- Automated static analysis tools evaluated as predictors of field failures
- Empirical Research of V&V in Concurrency:
  - TestCon method & Java PathFinder
  - Other:
    - Constrained Specification Based Testing
    - Atomizer
    - deadlock detection methods for Ada programs

Maximising Strategy

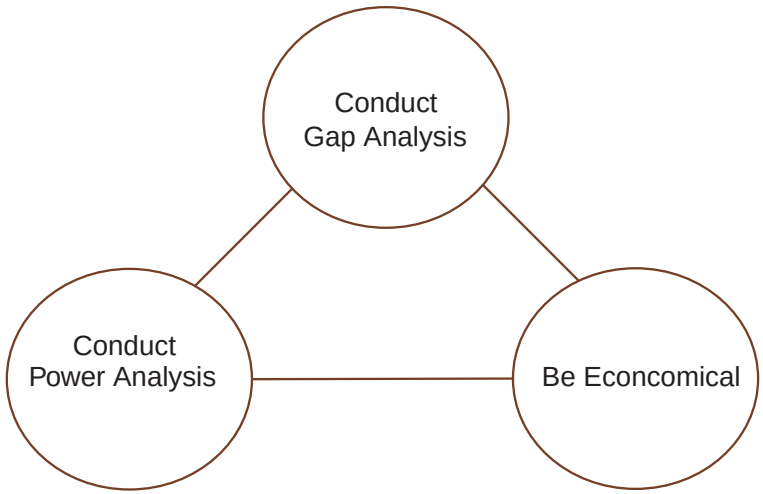
Gap Analysis

- Individual studies cannot produce generalisable results
  - Systematic literature review
  - Replication (close or differentiated)
  - Families of experiments (Basili et al.)



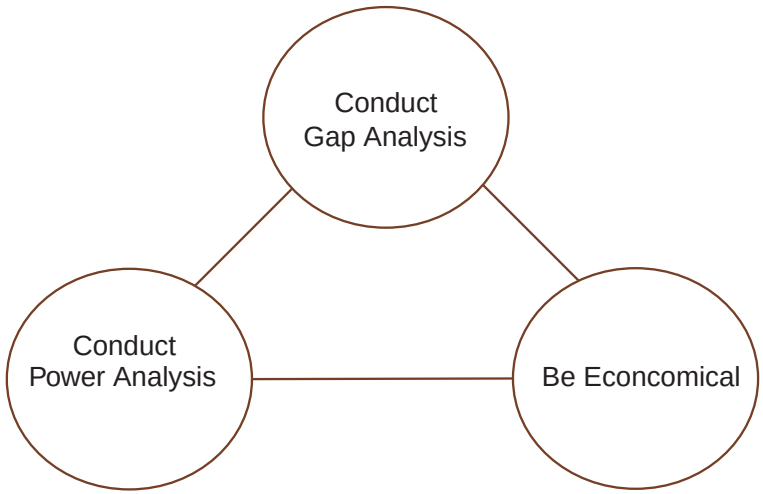
Power Analysis

- Balance Type II vs Type I errors
  - choose an appropriate α for your study (α> 0.05)
  - exploratory study requires higher power (correctly rejecting H_0)
- Use Cohen's sample size tables to determine number of participants
  - Information required: design, α, power and effect size



Being Economical

- Choose most economical experimental design (1-Way vs 2-Way ANOVA)
- Reuse experimental constructs
  - data collection forms
  - metrics (example: motivation)
  - benchmarks



The Experiment - Preparation

Goals, Hypothesis & Metrics

- Evaluating TestCon and combination of code inspection with automated static analysis
- Hypotheses

I. Cost-effectiveness improvement in terms of:

A. effectiveness (percentage of defects detected)
B. effectiveness (percentage of false positives detected)
C. efficiency (defects detected per hour)

II. Motivation and self-perceived mastery of the V&V technology vs effectiveness

Design & Power Analysis

- Design: 2-Way ANOVA
- Main effects: TestCon and Automated Static Analysis Tools
- Power analysis:
  - Cohens effect-size = 0.80
  - ANOVA with F-test (α=0.10 and Power =0.80)
  - minimum sample size is 4 per group

	TestCon - No	TestCon - Yes
Tools - No	(T1) Ad-hoc code inspection and code walkthrough	(T2) TestCon code inspection
Tools - Yes	(T3) Automated static analysis tools	(T4) TestCon and automated static analysis tools

Subjects & Objects

- Components and defects:
  - 4 concurrent components (Java)
  - 8 seeded and 2 prior defects
  - 6 of the 10 defects reported by the automated static analysis tools
  - 4 of the 10 defects only detectable by inspection of the code
- 16 COMP3402 students (90 min session):
  - training (15 minutes)
  - defect detection (60 minutes)
  - debriefing

Data Collection

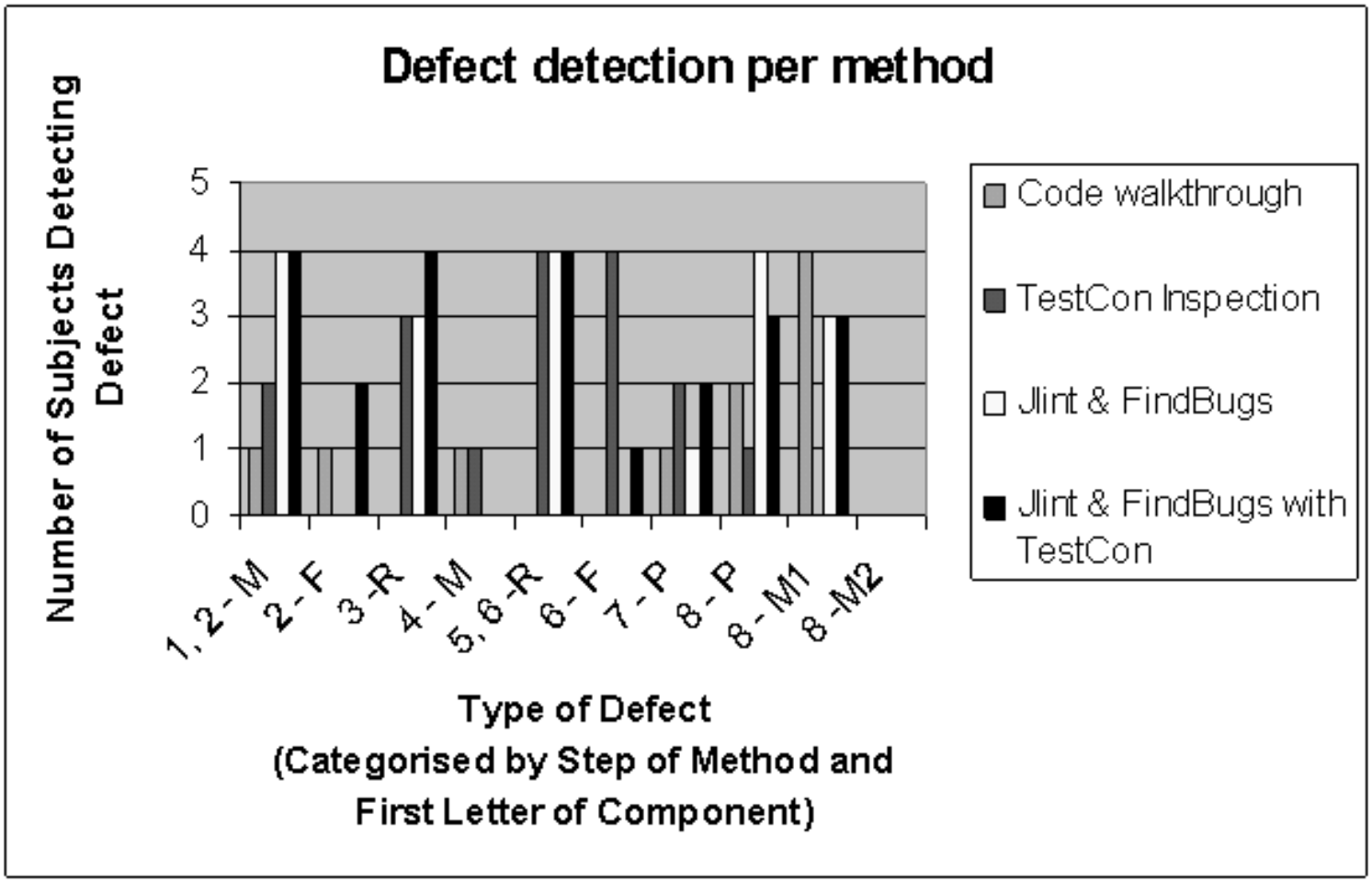
- Templates from Kamsties & Lott experimental package:
  - ☐ Information sheet prior to application of V&V technology
  - ☐ Defect detection report:
    - defects
    - time of detection
    - duration of TestCon step (if applicable)
  - ☐ Experimenter records time to apply V&V
  - ☐ After V&V appilcation participants report on:
    - perceived mastery
    - process conformance
    - motivation

The Experiment - Execution and Analysis

Execution

- Sample:
  - average of 3 years experience with Java
  - average of 4 courses with Java
  - no prior use of automated static analysis tools
- Preparation:
  - random assignment to one of four treatment groups
  - training (15 minutes)
  - overview of component interfaces (familiarisation prior to defect detection)

Analysis & Interpretation



I.A. Defect detection effectiveness

- Automated Static Analysis tools improve effectiveness of defect detection in terms of percentage of defects detected ($p<0.10$)

	TOOLUSED-N	TOOLUSED-Y	ROWS
TESTCON-N	.250 (.1732)	.450 (.1000)	.350 (.1690)
TESTCON-Y	.425 (.1500)	.550 (.2646)	.488 (.2100)
COLUMNS	.338 (.1768)	.500 (.1927)	.419 (.1974)

I.B. Defect detection effectiveness

- Both automated static analysis tools and TestCon improve the effectiveness of defect detection in terms of false positives detected ($p<0.10$)

	TOOLUSED-N	TOOLUSED-Y	ROWS
TESTCON-N	.7056 (.12620)	.3215 (.14232)	.5090 (.24423)
TESTCON-Y	.4214 (.22820)	.2468 (.13605)	.3341 (.19906)
COLUMNS	.5635 (.22849)	.2797 (.24423)	.4216 (.23342)

I.C. Defect detection efficiency

- Both automated static analysis tools and TestCon do not significantly improve the efficiency of defect detection

	TOOLUSED-N	TOOLUSED-Y	ROWS
TESTCON-N	3.6223 (2.93185)	7.1497 (2.65839)	5.3860 (3.20432)
TESTCON-Y	5.3034 (1.89792)	6.0953 (2.27549)	5.6994 (1.98545)
COLUMNS	4.4628 (2.45666)	6.6225 (2.35912)	5.5427 (2.58019)

Time of TestCon Steps

- Descriptive statistics of the mean time (std. dev.) in minutes of steps in the TestCon treatments
- Steps 6-8 take less time with tools.

Steps/Treatments	Step 2	Step 6	Step 7	Step 8
Treatment 2	3.19 (2.12)	3.37 (2.50)	4.83 (5.35)	4 (4.36)
Treatment 4	3.35 (1.68)	2.17 (0.99)	2.69 (1.46)	2.06 (1.01)

II. Correlation Testing

- Using Spearman rank-correlation test for ordinal values
- No correlation between motivation (r=0.344) and effectiveness
- No correlation for self-perceived mastery of V&V technology (r=0.059 Jlint and r=0.111 FB) and effectiveness

Threats to Validity and Conclusions

Threats to Validity

- Internal:
  - convenience sample (motivation did not influence performance) (-)
  - examination-like environment (-)
  - consistent training (+)
- External:
  - not random sample of software engineers (-)
  - small components (-)
  - limit of 60 minutes (-/+)
  - defects ensure that TestCon was evaluated (+)

Threats to Validity

- Conclusion:
  - percentage of defects and false positives and time repeatable and have high validity (+)
  - motivation and perceived mastery of the V&V technology do not have high validity (recorded after application like K&L) (-/+)
- Construct:
  - effectiveness and efficiency are common measures in evaluation of V&V technology (+)
  - motivation and perceived mastery have been used in prior studies (+)

Conclusions

- Automated static analysis tools
  - Improve defect detection effectiveness (percentage of defects detected)
  - Speed up TestCon
  - participants may become overly dependent on tools
- TestCon and Automated Static Analysis tools
  - Improve defect detection effectiveness by reducing the amount of false positives detected
  - Inform practitioner
  - Efficiency not improved, possibly due to size of components

Discussion of Maximising Strategy and Conclusion

Maximising Strategy - Advantages

- Study is not performed in isolation
  - context of past and future related studies
- Guide research and encourage replication
  - overcome threats to validity of past studies
- Information and analysis gained despite small size of study
  - dependence on tools can occur despite process conformance
  - effectiveness is improved with respect to false positives
- Effect size for percentage of defects detected that were false positives was accurate

Maximising Strategy - Drawbacks

- Lower statistical significance
- Effect size for percentage of defects detected was actually smaller than expected (more subjects required for future studies with the same significance)
- Limited generalisation
- Possible effects on results (efficiency)

Conclusion

- Empirical studies are costly therefore maximising information gained is important
- Maximise information gained through:
  - gap analysis
  - power analysis
  - being economical (reducing preparation and execution costs)
- Study supports combination of V&V technologies for concurrent components with empirical evidence
- Overcome limitations of single study to encourage empiricism in new contexts

Questions & Answers