Blekinge Institute of Technology

Claes Wohlin

Software Inspections: A Series of Reading Experiments

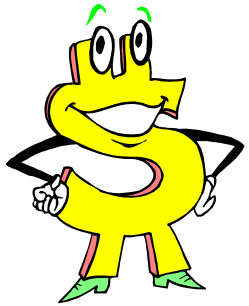
Claes Wohlin

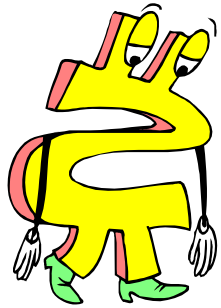
Blekinge Institute of Technology

Outline of Presentation

- Background
- Usage-Based Reading
- A series of experiments
  - Common information
  - Three experiments
- Conclusions

Moving the User Perspective





Analysis

System Test

Design

Function test

Implementation

Inspections

Inspections contain several steps:

- **Overview**
- **Individual preparation**
- **Team meeting**
- **Re-work**

In the individual preparation different techniques may be used to support the reviewer.

Techniques in the Individual Preparation

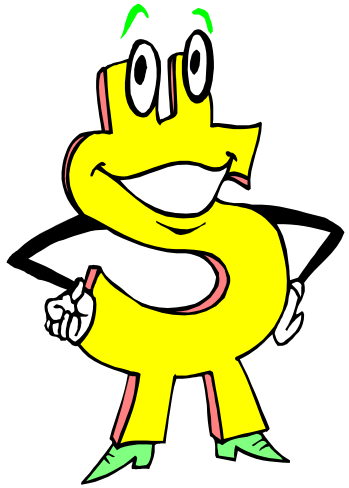
- Ad hoc
- Checklists
- Perspective-Based Reading
- Defect-Based Reading

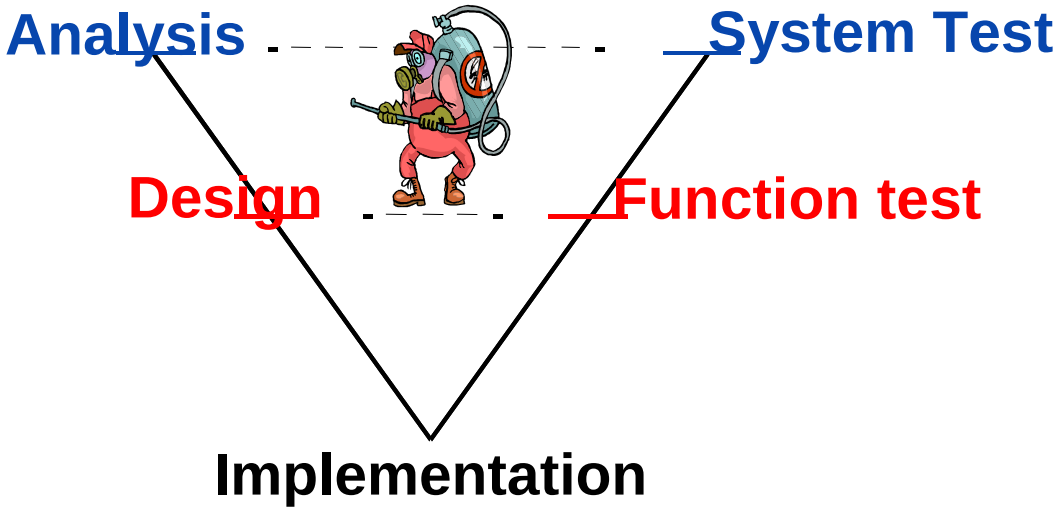
Our new proposal:

- Usage-Based Reading

Research Inspiration

- Use cases have been proposed as a user view in object-orientation
- Statistical usage testing has been proposed as part of a method denoted Cleanroom Software Engineering





Usage-Based Reading (UBR) Research

- Prioritized use cases
- Focus on user view
- Experiments
  - 1st Prioritized vs. random
  - 2nd UBR vs. checklist
  - 3rd Active vs. passive

Usage-Based Reading

- Inspect from a user perspective
- Let use cases drive the inspection
- Prioritise use cases from a user perspective, i.e. create a focus on critical faults from a usage point of view. UBR is defined as use case driven reading with prioritised use cases.

Experiment Context

- **Verification and Validation courses at Lund University and Blekinge Institute of Technology**
- **Experiment package**
  - **Taxi Management System**
  - **Textual requirements document**
  - **Use case document (24 use cases)**
  - **Design document**

Fault Classification

- **Class A: Crucial for the user**
- **Class B: Important for the user**
- **Class C: Irritating for the user**

Other Concerns

- **Experience questionnaire with seven questions capturing the background of the subjects**
- **Different people have contributed in the development of the system and the use cases. Moreover, classification of faults, and design and analysis of the experiments have been conducted by several people.**

Main Research Questions

- **Experiment 1: Is the prioritisation of use cases better than a random order?**
- **Experiment 2: Is Usage-Based Reading better than using a checklist?**
- **Experiment 3: Is it better to develop the use cases than using existing use cases?**

Common Design

- **Subjects: students at third or fourth year at the universities**
- **Control variable: student experience**
- **Dependent variable: time spent and faults**

Evaluation Variables

- **Effectiveness**
  - **Number of faults found out of the total number of faults**
- **Efficiency**
  - **Faults found per time unit**

Experiment 1

- **UBR with prioritised use cases vs. a random order of use cases**

General question:

- **Is UBR better with respect to effectiveness and efficiency than inspections with use cases in random order?**

Presentation Experiment 1

- Design
- Operation
- Descriptive analysis
- Statistical analysis
- Conclusions

Design 1(2)

Faults:

- **37 faults (A: 13; B: 13 and C: 11)**
- **17 were found during development**
- **8 were seeded**
- **12 new faults in the experiment**
- **Syntax faults are not calculated**

Design 2(2)

- **Control variable: experience of subjects, although no significant difference between subjects.**
- **27 students divided into two groups (14 respectively 13 students).**

Hypotheses

- Reviewers using UBR are more efficient (overall, for faults of type A, and for faults of type A+B)
- Reviewers using UBR are more effective (overall, for faults of type A, and for faults of type A+B)
- Reviewers using UBR defect different faults

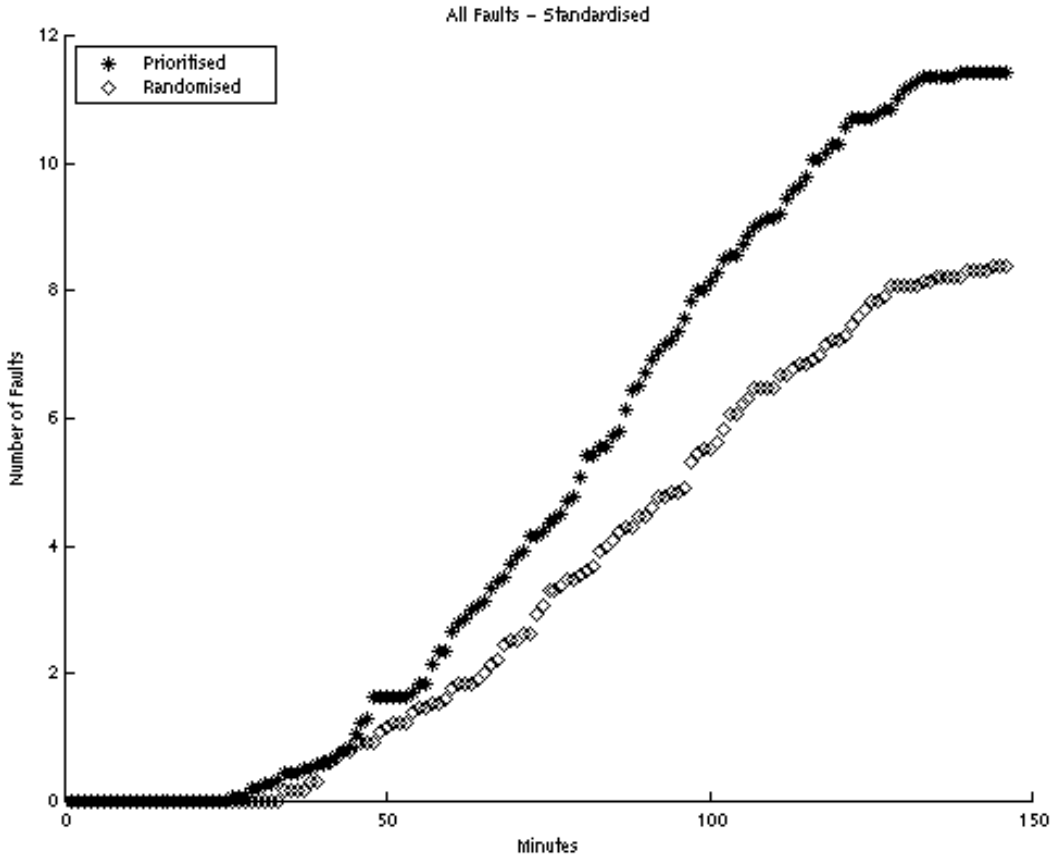
Operation

- Fall of 2000
- Students were familiar with the type of system from a prior course
- Mandatory part of course, although anonymity guaranteed
- Total time for experiment: 2.5 hours

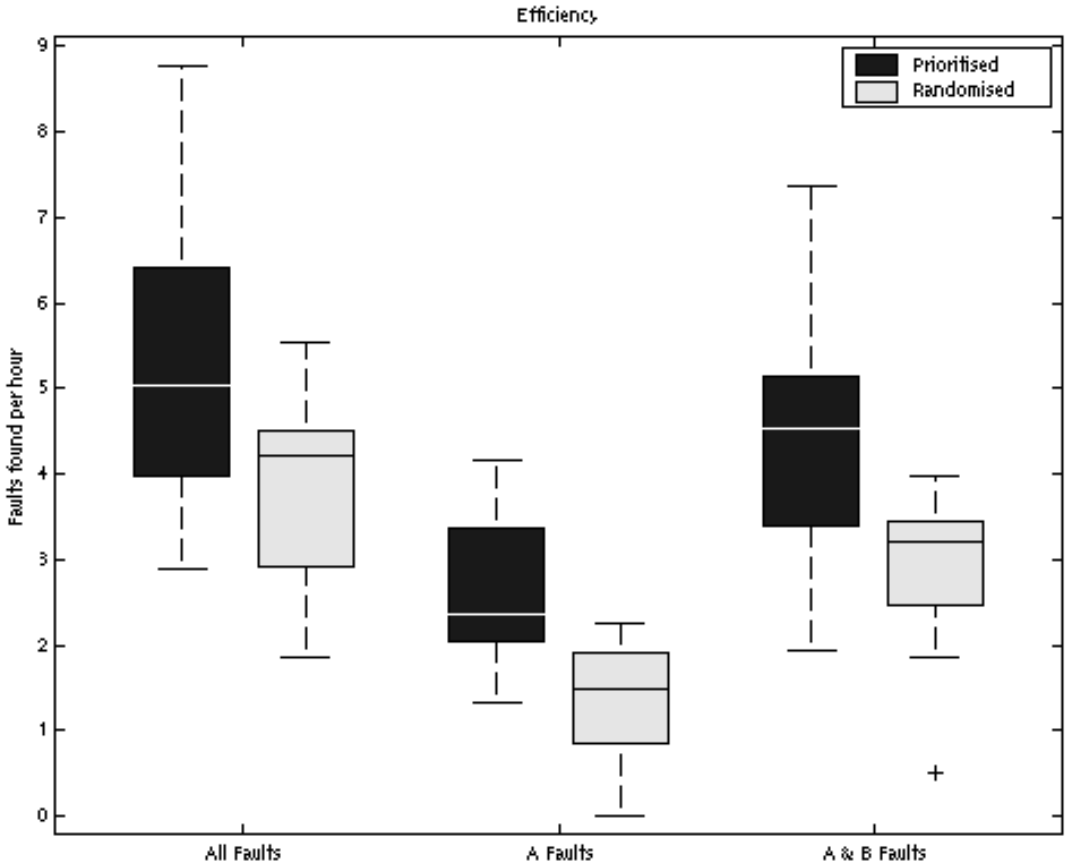
Time Spent in Experiment

	Mean (minutes)		Standard Deviation	
	Prioritised	**Randomised**	**Prioritised**	**Randomised**
Preparation	31.1	43.6	8.1	18.5
Inspection	99.1	87.2	11.5	16.9
Total	129.8	130.8	9.8	10.2

Cumulative Number of Faults



Box Plots of the Efficiency



Test of Hypotheses

Mann-Whitney test since the data are not normally distributed.

	Efficiency (P value)	Effectiveness (P value)
All Faults	0.0440	0.0652
Class A Faults	0.0004	0.0017
Class A & B Faults	0.0049	0.0045

Significance level 0.95: Support for efficiency hypotheses and for effectiveness with respect to Class A and Class A+B In addition, the groups detect different faults.

Conclusions

- **UBR reviewers are more efficient**
- **UBR reviewers are more effective for crucial and important faults from the user perspective**
- **UBR reviewers find different faults than those using use cases in random order.**

Experiment 2

- **UBR vs. Inspections using a checklist**

General question:

- **Is UBR better with respect to effectiveness and efficiency than checklist-based reading?**

Presentation Experiment 2

- Design
- Operation
- Descriptive analysis
- Statistical analysis
- Conclusions

Design 1(2)

Faults:

- **38 faults (A: 13; B: 14 and C: 11)**
- **28 were found during development or in inspections and test**
- **8 were seeded**
- **2 new faults in the experiment**
- **Syntax faults are not calculated**

Design 2(2)

- Control variable: experience of subjects. This resulted in three groups from which the students were randomised into two groups.
- 27 students divided into two groups (12 respectively 11 students).
- Independent variable
  - checklist
  - use cases in prioritised order

Hypotheses

- The reviewers applying UBR are more *efficient*
- The reviewers applying UBR are more *effective* than the reviewers using a checklist
- The reviewers applying UBR finds different faults
- The reviewers applying UBR are more effective and efficient as a team as well

Operation 1(2)

- Spring of 2001
- The experiment was held over 2 days
- Mandatory part of course, although anonymity guaranteed
- Schedule according to the schedule on the next slide

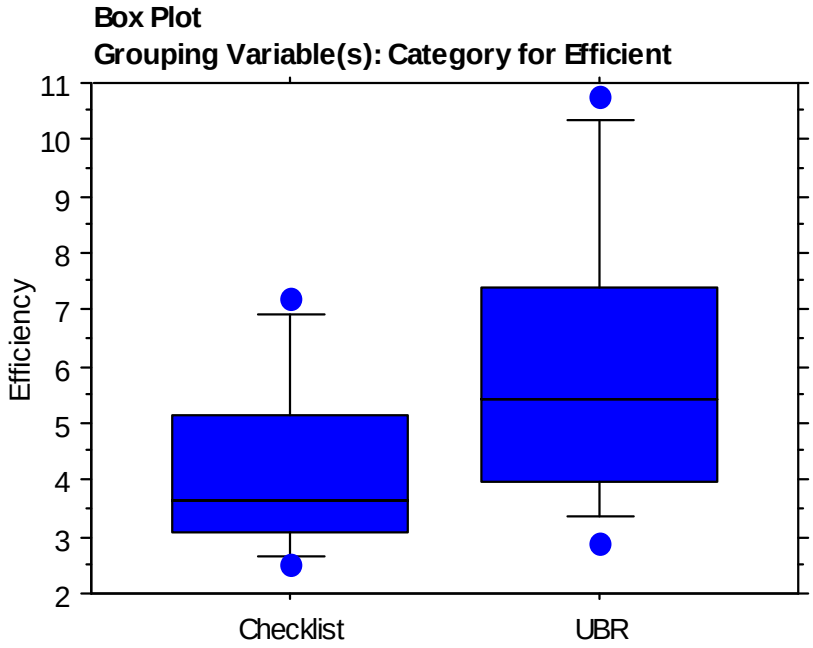
Operation 2(2)

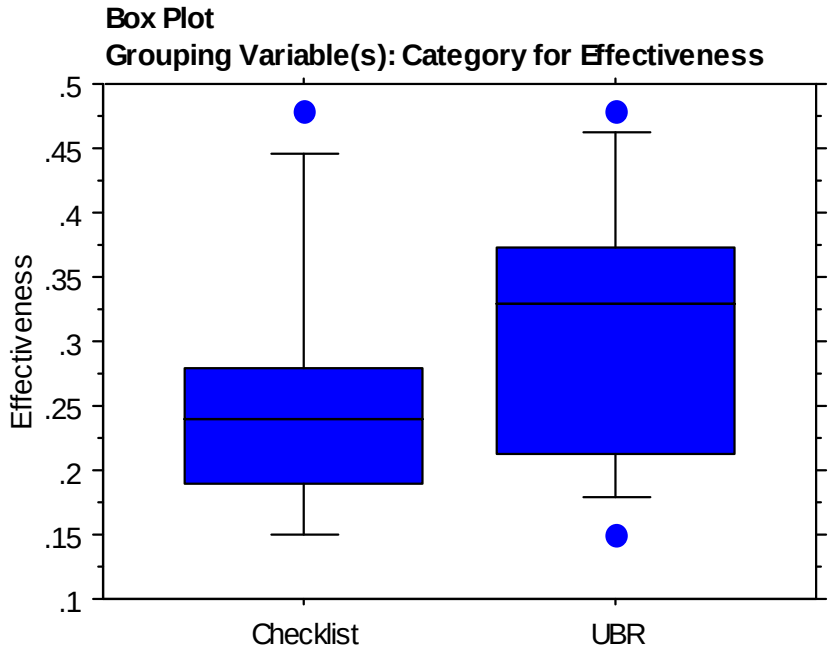
	CBR group	UBR group
Day 1 (1.15 p.m. - 2.00 p.m.)	General introduction to the Taxi Management System	
Day 1 (2.15 p.m. - 3.00 p.m.)	Introduction to CBR	Introduction to UBR
Day 2 (9.15 a.m. - 12.00 p.m.)	Inspection Experiment	
Day 2 (1.15 p.m. - 2.00 p.m.)	Introduction to UBR and follow-up discussion	Introduction to CBR and follow-up discussion

Comparison: Efficiency and Effectiveness

	More faults found	More Unique Faults
All Faults	21.1%(UBR)	10.0%(UBR)
Class A Faults	75.1%(UBR)	18.2%(UBR)
Class B Faults	27.7%(UBR)	20.0%(UBR)
Class C Faults	*62.5%(CBR)*	*12.5%(CBR)*
Class A&B Faults	50.5%(UBR)	19.0%(UBR)

Box Plot: All faults





Statistical Test

- **Mann Whitney (data are not normally distributed)**

	Efficiency (P value)	Effectiveness (P value)
All Faults	0.0423	*0.1029*
Class A Faults	0.0127	0.0364
Class A & B Faults	0.0164	0.0312
Class B Faults	*0.1481*	*0.1754*
Class C Faults	0.2679	0.1481

Conclusions

- UBR reviewers are more efficient (all, type A and type A+B)
- UBR reviewers are more effective for type A and type A+B
- UBR reviewers detect different faults than those using a checklist
- UBR teams are better than CBR or mixed teams

Experiment 3

- **UBR with utilizing use cases vs. Inspections with developing use cases.**

General question:

- **Is UBR with utilizing use cases better with respect to effectiveness and efficiency than if the reviewer develops use cases?**

Presentation Experiment 3

- Design
- Operation
- Descriptive analysis
- Statistical analysis
- Conclusions

Design

Faults:

- **38 faults (A: 13; B: 14 and C: 11)**
- **Control variable**
  - **Experience**
  - **Student characterisation form => two groups of students**
  - **Randomised within the two groups to form the red and green groups respectively**
- **Independent variable: develop vs. utilise**

Hypotheses

- The reviewers developing use cases are equally *efficient* as the reviewers utilizing use cases

- The reviewers developing use cases are equally *effective* as the reviewers utilizing use cases

Operation

- Fall of 2001
- The experiment was held over 2 days (similar to experiment 2)
- However, an inspection meeting was added.
- Mandatory part of course, although anonymity guaranteed

Data (average) – Individual

	Green	Red
Number of Faults Found	12.5	12.1
Preparation Time	32	28
Inspection Time	105	137
Total Time	137	165
Efficiency (faults found per hour)	5.9	4.6
Effectiveness (found faults / total)	0.33	0.32

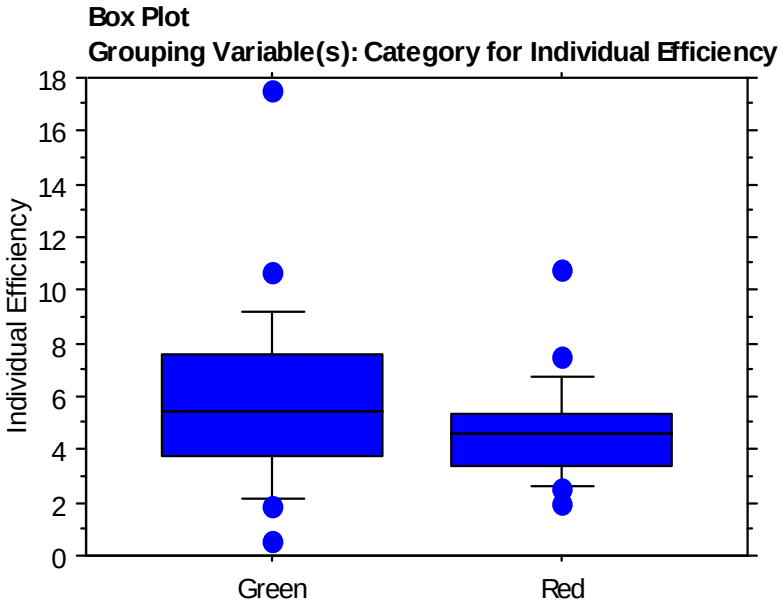
- **38 faults in the document**
- **Red – develops and Green - utilizes**

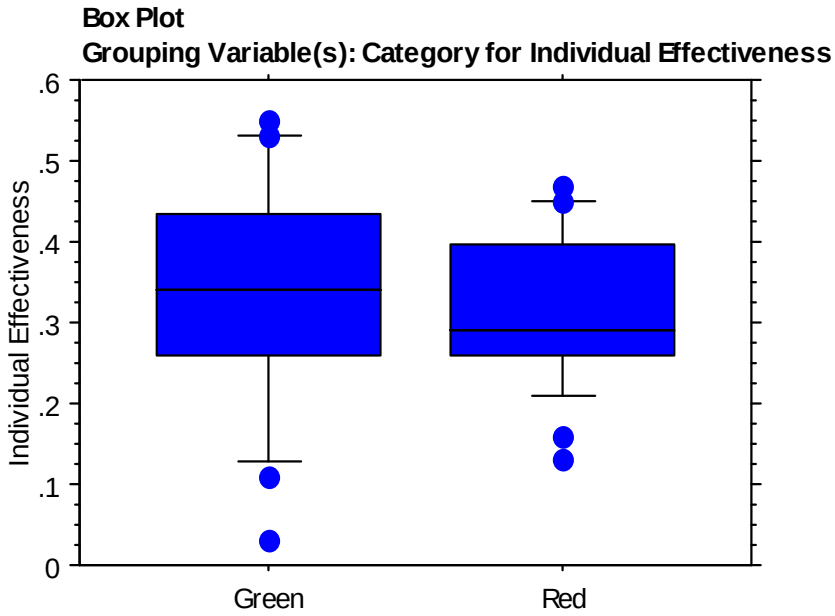
Data (average) – Total

	Green	Red
Number of Faults Found	21	21
Meeting Time	68	58
Efficiency (Total)	2,76	2,25
Effectiveness	0,56	0,56

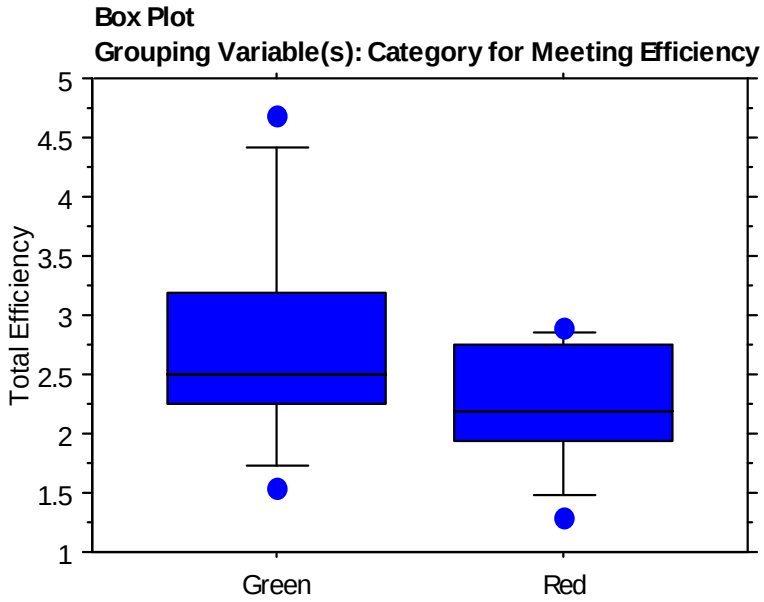
- **No new faults were found during the meeting**

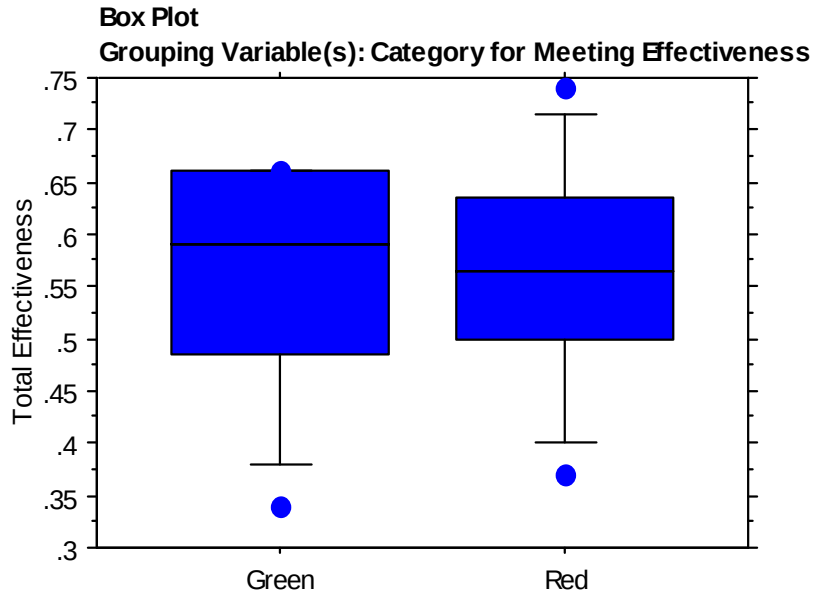
Box Plot (Individual)





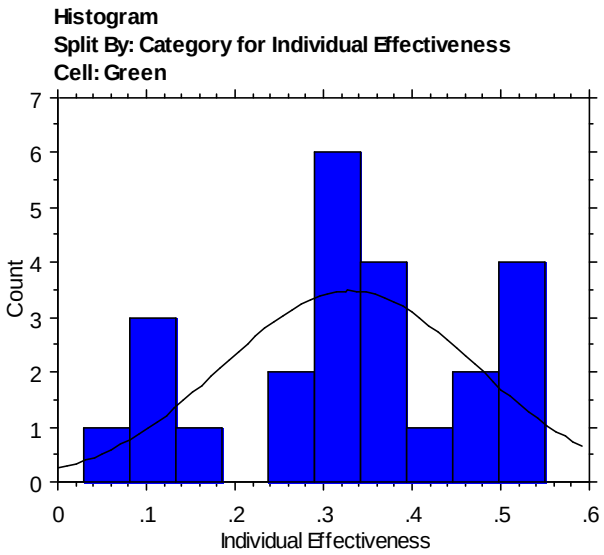
Box Plot (Total)





Statistical Test

- **ANOVA**
  - Assumes a normal distribution and equal variances
  - Normal probability plots
  - Kolmogorov-Smirnov test
  - Residual plots
- **Mann-Whitney**
  - non-parametric test due to differences in variances and the data are not normally distributed
  - Significance level = 0.05



Result (non-parametric analysis)

- **Individual**
  - **Efficiency: p-value = 0.13**
  - **Effectiveness: p-value = 0.59**
- **Total**
  - **Efficiency: p-value = 0.40**
  - **Effectiveness: p-value = 0.95**

No significant differences.

Discussion: Experiment 3

- **Result**
  - **No statistical significant results**
  - **Interpretation, explanation?**
- **Discussion**
  - **Include time for developing the use cases (green)?**
  - **Did the groups find different types of faults?**
  - **Why were not any new faults found during meeting?**

Additional experiment

Experiment 3 was also run at another site and when combining the results, a significant difference emerged. Reviewers utilizing pre-developed use cases were more efficient.

General Conclusions

- **UBR is better than checklist-based reading**
- **UBR with prioritised use cases is better than having random order use cases**
- **Utilizing or developing use cases is still an open issue, although it is leaning towards utilizing**

Conclusions: Experiment Series

The series of experiments shows that it is possible to create different experiments evaluating one aspect at the time.

Thus, series of experiments is an important tool for evaluating different methods and techniques.