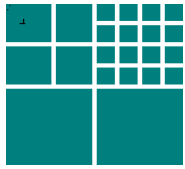


# Experimentation in Software Engineering: Reading Studies II



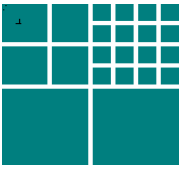
## Scenario-Based Reading Definition

---

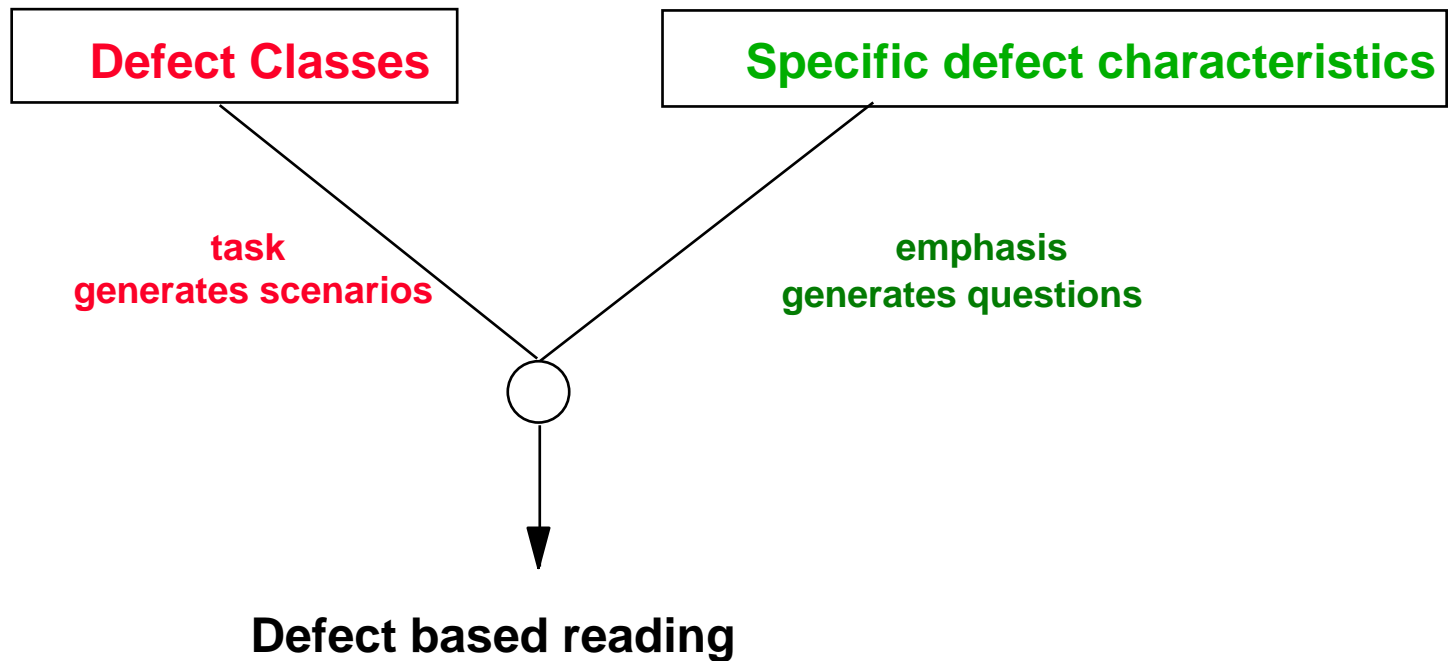


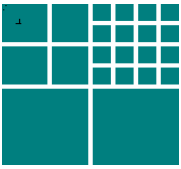
Another approach, **defect-based reading** focuses on procedures used for identifying different types of defects, e.g., data type consistencies, incorrect functionalities, ambiguities or missing functionality, ... and the emphasis is on questions associated with helping identify these types of defects.

The technique we are studying is associated with reading requirements/specification documents in the Software Cost Reduction (SCR) Notation, a state machine transition notation developed by Dave Parnas.



# Reading for Analysis: Defect-Based Reading Definition





# Reading for Analysis: Blocked Subject-Project Study

---



## Defect-Based Reading

### Study Goal:

Analyze defect-based reading, ad-hoc reading and check-list based reading to evaluate and compare them

with respect to their effect on fault detection effectiveness in the context of an inspection team

from the viewpoint of quality assurance

### Environment:

University of Maryland graduate courses

Requirements documents written in SCR notation

Water Level Monitoring System, Cruise Control System

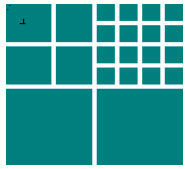
### Experimental design:

Blocked subject-project: Partial factorial design

Replicated twice

Subjects: 48 subjects in total

Note: 1A – first replication, team A



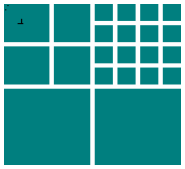
# Defect-Based Reading



## Round/Specification

**Detection  
Method**

	Round 1		Round 2	
	WLMS	CRUISE	WLMS	CRUISE
ad hoc	1B, 1D, 1G, 1H, 2A	1A, 1C, 1E, 1F, 2D	1A	1D, 2B
checklist	2B	2E, 2G	1E, 2D, 2G	1B, 1H
scenarios	2C, 2F	2H	1F, 1C, 2E, 2H	1G, 2A, 2C, 2F



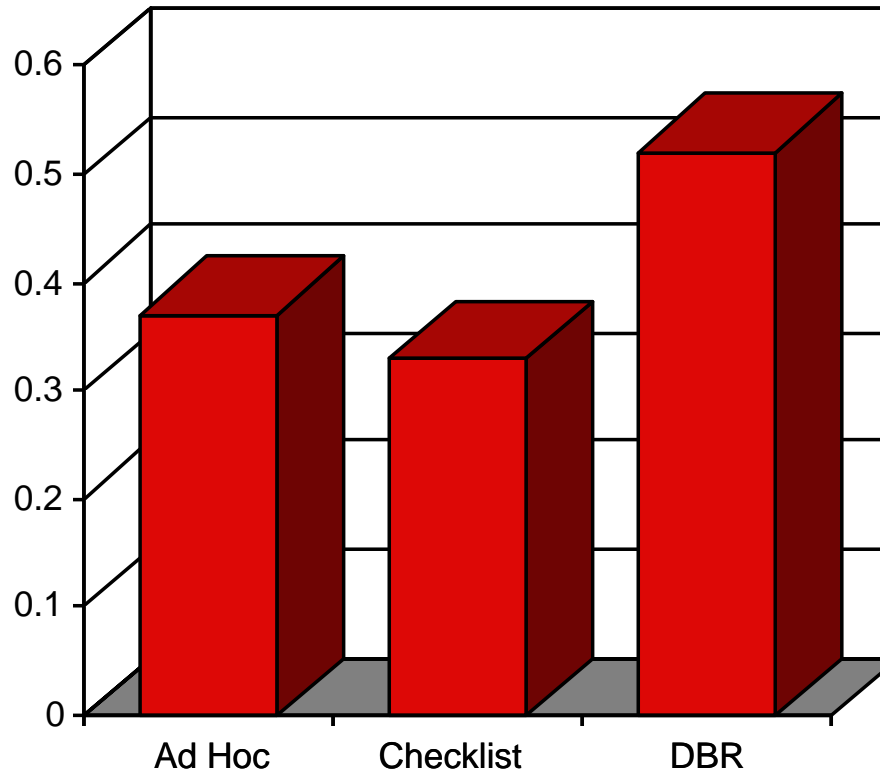
# Reading for Analysis: Defect-Based Reading Experiment

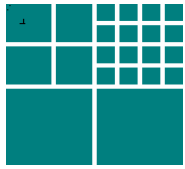


Goal of Defect-Based Reading (DBR):  
detect defects in a requirements document  
focus on [defect classes](#)

Controlled experiment run twice with UMD graduate students:

**Team  
Detection  
Rate**





# Reading for Analysis: Blocked Subject Project Study

---



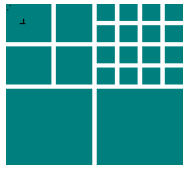
## Defect-Based Reading

### Major Results

**Scenario** readers performed better than Ad Hoc and Checklist  
Readers improvement of about 35%

**Scenarios** helped reviewers focus on specific fault classes but were  
no less effective at detecting other faults

**Checklist reading** was no more effective than Ad Hoc reading



# The Experimental Discipline

---



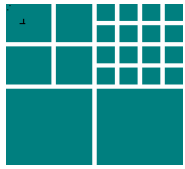
## Experimental Classifications

### Level of variable relationship

**Descriptive:** there may be patterns in the data but the relationship among the variables has not been examined

**Correlational:** the variation in the dependent variable(s) is related to the variation of the independent variable (s)

**Cause-effect:** the treatment variable(s) is the only possible cause of variation in the dependent variable(s)



# The Experimental Discipline

---



## Experimental Classifications

### Experience of Subjects

**novice:** students or individuals not experienced in domain

**experts:** practitioners or people with experience in domain

### Experimental Setting

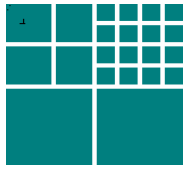
**In vivo:** in the field under normal conditions

**In vitro:** in the laboratory under controlled conditions

### Type of Study

**Experiment:** at least one treatment or controlled variable

**Observational study:** no treatment or controlled variables



## Experimental Classifications

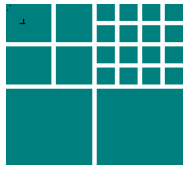
### Types of Analysis

#### **Quantitative Analysis**

- obtrusive controlled measurement
- objective
- verification oriented

#### **Qualitative Analysis**

- naturalistic and uncontrolled observation
- subjective
- discovery oriented



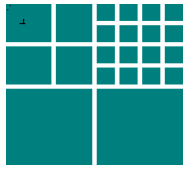
## Experimental Classifications

### Study

- an act to discover something unknown or of testing a hypothesis
- can include all forms of quantitative and qualitative analysis

Studies can be

- **experimental**
  - driven by hypotheses; quantitative analysis
  - controlled experiments
  - quasi-experiments or pre-experimental designs
- **observational**
  - driven by understanding; qualitative analysis dominates
  - qualitative/quantitative study
  - pure qualitative study



# The Experimental Discipline

---



## Experimental Study Classifications

**Experiments** can be

- controlled experiments
- quasi-experiments or pre-experimental designs

**Controlled experiments**, typically:

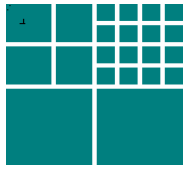
- small object of study
- in vitro
- a mix of both novices (mostly) and expert treatments

Sometimes, novice subjects used to “debug” the experimental design

**Quasi-experiments** or **Pre-experimental design**, typically:

- large projects
- in vivo
- with experts

These latter experiments tend to involve a qualitative analysis component, including at least some form of interviewing



# Experimental and Quasi-Experimental Designs

---



Experimentation is not a panacea, but rather the only available route to cumulative progress

There are a large variety of experimental and quasi-experimental designs

These are represented in what follows, using the notation:

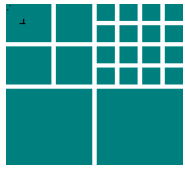
Let X represent the exposure of a group to an experimental variable or event, the effects of which are to be measured

Let O refer to some process of observation or measurement

Assume the X's and O's in the same line are given to the same specific persons

Let R represent the random assignment to separate groups

- Campbell & Stanley, Experimental and Quasi-experimental Designs for Research

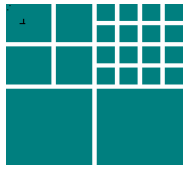


# Factors Jeopardizing Validity

---



- There are several factors that can jeopardize the validity of an experimental design
- They can be broken into **internal** and **external** validity
- Internal validity is the basic minimum without which an experiment is uninterpretable
  - Did in fact the experimental treatments make any difference in this specific experimental instance?
- External validity deals with the issue of generalizability
  - To what populations, settings, treatment variables, and measurement variables can this effect be generalized?
    - Campbell & Stanley, Experimental and Quasi-experimental Designs for Research



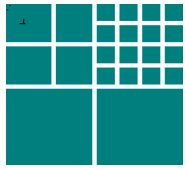
# Internal Validity

---



Eight different classes of extraneous variables, which, if not controlled in the experimental design, might produce effects confounded with the effect of the experimental stimulus.

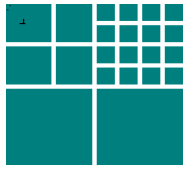
- History - the specific events occurring between the first and second measurement in addition to the experimental value, creating rival hypotheses (O1 X O2)
- Maturation - processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, hungrier, more tired, learning, etc.
- Testing - the effects of taking a first test upon the scores of a second testing, i.e., know how to answer
- Instrumentation - changes in the calibration of a measuring instrument or changes in the observers or scorers used, may produce changes in the obtained measurements.
  - Campbell & Stanley, Experimental and Quasi-experimental Designs for Research



# Internal Validity

---

- Statistical Regression - operating where groups have been selected on the basis of their extreme scores, i.e., tendency toward the mean
- Selection - biases resulting in differential selection of respondents for the comparison groups (X O1, O2), e.g., volunteers are more enthusiastic
- Experimental Mortality - differential loss of respondents from the comparison groups
- Selection-Maturation Interaction, etc. - any of the extraneous variables can have a combined effect that can be mistaken for the effect of the experimental variable
  - Campbell & Stanley, Experimental and Quasi-experimental Designs for Research



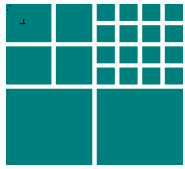
# External Validity

---



The factors jeopardizing external validity or representativeness are:

- **Testing and X: Reactive/Interaction Effect of Testing** - a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable, thus making the results obtained for a pre-tested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected
- **Selection and X: Interaction** effects of selection biases and the experimental variable, i.e. need to make sure the population is representative
  - Campbell & Stanley, Experimental and Quasi-experimental Designs for Research

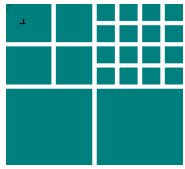


# External Validity

---



- Reactive Effects of Experimental Arrangements - preclude generalization about the effect of the experimental variable upon persons being exposed to it in non-experimental settings, e.g., in vitro does not always generalize to in vivo as the setting is different
- Multi-Treatment Interference - likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable. This is a particular problem for one-group designs of type 8 or 9.
  - Campbell & Stanley, Experimental and Quasi-experimental Designs for Research



# Pre-Experimental Designs

---

## Design 1: The One Shot Case Study

X O

Absence of control, almost no scientific value, opportunity for qualitative analysis or technique evolution

Rival Hypotheses: history, maturation, selection, mortality,

## Design 2: The One Group Pretest Posttest design

O<sub>1</sub> X O<sub>2</sub>

Rival hypotheses: history, maturation, testing, instrumentation, statistical regression (?),

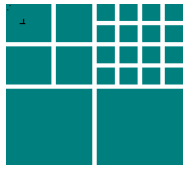
## Design 3: The Static Group Comparison

X O<sub>1</sub>

O<sub>2</sub>

Rival hypotheses: selection, mortality,

- Campbell & Stanley,
- Experimental and Quasi-experimental Designs for Research



# True Experimental Designs

---

## Design 4: The pretest post test Control Group design

R O<sub>1</sub> X O<sub>2</sub>  
R O<sub>3</sub> O<sub>4</sub>

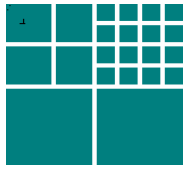
## Design 5: The Solomon Four group design

R O<sub>1</sub> X O<sub>2</sub>  
R O<sub>3</sub> O<sub>4</sub>  
R X O<sub>5</sub>  
R O<sub>6</sub>

## Design 6: Posttest Only Control Group Design

R X O<sub>1</sub>  
R O<sub>2</sub>

– Campbell & Stanley, Experimental and Quasi-experimental Designs for Research



# True Experimental Designs

---

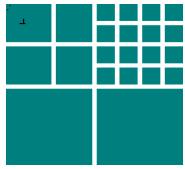
## Factorial Designs: Several treatments (ala Design 6)

R  $X_1$   $O_1$   
R  $X_2$   $O_2$   
R  $X_3$   $O_3$   
...  
R  $X_n$   $O_n$

Can be done with Design 4 and 5 also

Can be done with a control group as well

- Campbell & Stanley, Experimental and Quasi-experimental Designs for Research



# Quasi-Experimental Designs

---

When the experimenter lacks full control over the scheduling of experimental stimuli, something like an experimental design can be introduced

## Time Series Design

$O_1 O_2 O_3 O_4 X O_5 O_6 O_7 O_8$

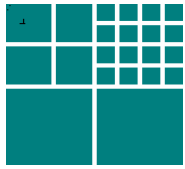
## Equivalent Time Samples Design

$X_1 O_1, X_2 O_2, X_1 O_3, X_2 O_4, \dots$

## Non-Equivalent Control Group Design

$O X O$   
 $O O$

Campbell & Stanley,  
Experimental and Quasi-experimental Designs for Research



# Questions about the results of a study

---

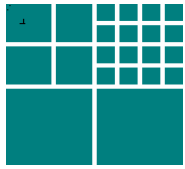


We have seen several experiments, all building on the idea of what are the best way to analyze software documents.

But were these good experiments? What kinds of experimental designs were used? Are they reasonable design? How valid are the results?

Did a good reading method/technique actually *cause* there to be less failures? Do the results generalize to all practitioners? Does it represent a real environment?

Does the data, without some other form of analysis provide the full understanding, insights into what happened? What variables were controlled for and what were not?



# Blocked Subject Project Study



## Testing/Reading Strategies Comparison

### Goals:

Analyze code reading, functional testing and structural testing to evaluate and compare them with respect to their effect on fault detection effectiveness, fault detection cost and classes of faults detected from the viewpoint of quality assurance

### Environment:

NASA/CSC and the University of Maryland  
Text formatter, plotter, abstract data type, database  
Seeded with software faults (9, 6, 7, 12)  
145 - 365 LOC

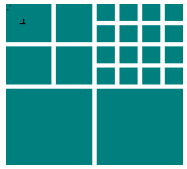
### Experimental design:

**Factorial Designs: Several treatments**

R  $X_1$   $O_1$

R  $X_2$   $O_2$

32 NASA/CSC Subjects



# Blocked Subject Project Study Testing/Reading Strategies Comparison

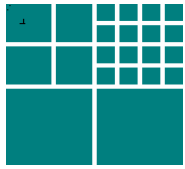


## Fractional Factorial Design

		<u>Code Reading</u>			<u>Functional Testing</u>			<u>Structural Testing</u>		
		P1	P2	P3	P1	P2	P3	P1	P2	P3
Advanced Subjects	S1			X		X		X		
	S2		X		X					X
	:									
	S8	X					X		X	
Intermediate Subjects	S9			X		X		X		
	S10		X		X					X
	:									
	S19	X					X		X	
Junior Subjects	S20			X		X		X		
	S21		X		X					X
	:									
	S32	X					X		X	

**Blocking according to experience level and program tested  
Each subject uses each technique and tests each program**

**NASA/CSC**



# What is wrong with this design?

---

## Internal issues?

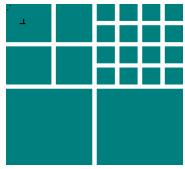
History, Maturation, Testing, Instrumentation, Statistical regression, Selection, Mortality, Selection-Maturation Interaction

Most look ok.

## External issues?

Testing and X, Selection and X, Reactive Effects of Experimental Arrangement, Multi-treatment Interference

Reactive Effects of Experimental Arrangement - they will be using the reading techniques on real projects, and as we saw, they did not apply them when they knew they were going to test.



# Replicated Project Study



## Cleanroom Study

### Study Goal:

Analyze the Cleanroom process in order to evaluate and compare it to a non-Cleanroom process with respect to the effects on the process, product and developers from the viewpoint of quality assurance

### Environment:

University of Maryland  
Electronic message system, ~ 1500 LOC

### Experimental design:

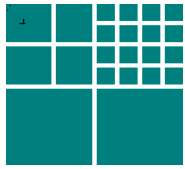
Replicated project, 15 three-person teams (10 used Cleanroom)

#### Posttest Only Control Group Design

R Cleanroom  $O_1$

R Non-Cleanroom  $O_2$

**Observations:** Attitude survey, On-line activity, Testing results  
3 to 5 test submissions



# What is wrong with this design?

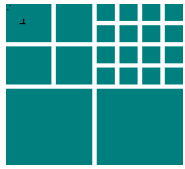
---

## **Internal issues?**

History, Maturation, Testing, Instrumentation, Statistical regression,  
Selection, Mortality, Selection-Maturation Interaction

## **External issues?**

Testing and X, Selection and X, Reactive Effects of Experimental  
Arrangement, Multi-treatment Interference



# Single Project Study



## Cleanroom in the SEL

### Study Goal:

Analyze the Cleanroom process  
in order to evaluate and compare it to a standard SEL development  
process  
with respect to the effects on the effort distribution, cost, and  
reliability  
from the viewpoint of quality assurance

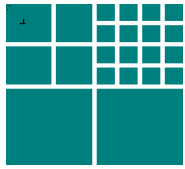
### Environment:

NASA/ SEL  
40 KLOC Ground Support System

### Experimental design:

**Time Series Design:**  $O_1 O_2 O_3 O_4 X O_5 O_6 O_7 O_8$

Data collected: effort distribution, change profile, productivity,  
level of rework, impact of spec changes, error  
rate, error distribution, error source



# What is wrong with this design?

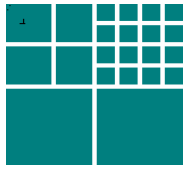
---

## **Internal issues?**

History, Maturation, Testing, Instrumentation, Statistical regression,  
Selection, Mortality, Selection-Maturation Interaction

## **External issues?**

Testing and X, Selection and X, Reactive Effects of Experimental  
Arrangement, Multi-treatment Interference



## Cleanroom in the SEL

### Study Goal:

Analyze the Cleanroom process

in order to evaluate and compare it to a standard SEL development process with respect to the effects on the effort distribution, cost, and reliability

### Environment:

Project 2: 22 KLOC Flight Dynamics System (in-house)

Project 3: 160 KLOC Flight Dynamics System (contractor)

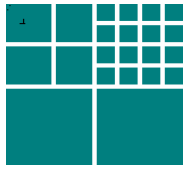
Project 4: 140 KLOC Flight Dynamics System (contractor)

### Experimental design:

#### Equivalent Time Samples Design:

$X_1 O_1, X_2 O_2, X_1 O_3, X_2 O_4, \dots$

Data collected: effort distribution, change profile, productivity, level of rework, impact of spec changes, error rate, error distribution, error source



# What is wrong with this design?

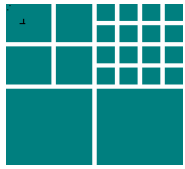
---

## **Internal issues?**

History, Maturation, Testing, Instrumentation, Statistical regression,  
Selection, Mortality, Selection-Maturation Interaction

## **External issues?**

Testing and X, Selection and X, Reactive Effects of Experimental  
Arrangement, Multi-treatment Interference



# Reading for Analysis: Blocked Subject Project Study

---



## Perspective-Based Reading

### Study Goal:

Analyze perspective-based reading, NASA's current reading technique to evaluate and compare them with respect to their effect on fault detection effectiveness in the context of an inspection team from the viewpoint of quality assurance

### Environment:

NASA/CSC SEL Environment

Requirements documents:

All documents seeded with known defects

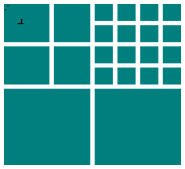
### Experimental design:

Blocked subject-project: Partial factorial design

R O<sub>1</sub> O<sub>3</sub> X<sub>1</sub> O<sub>5</sub> O<sub>7</sub>

R O<sub>2</sub> O<sub>4</sub> X<sub>2</sub> O<sub>6</sub> O<sub>8</sub>

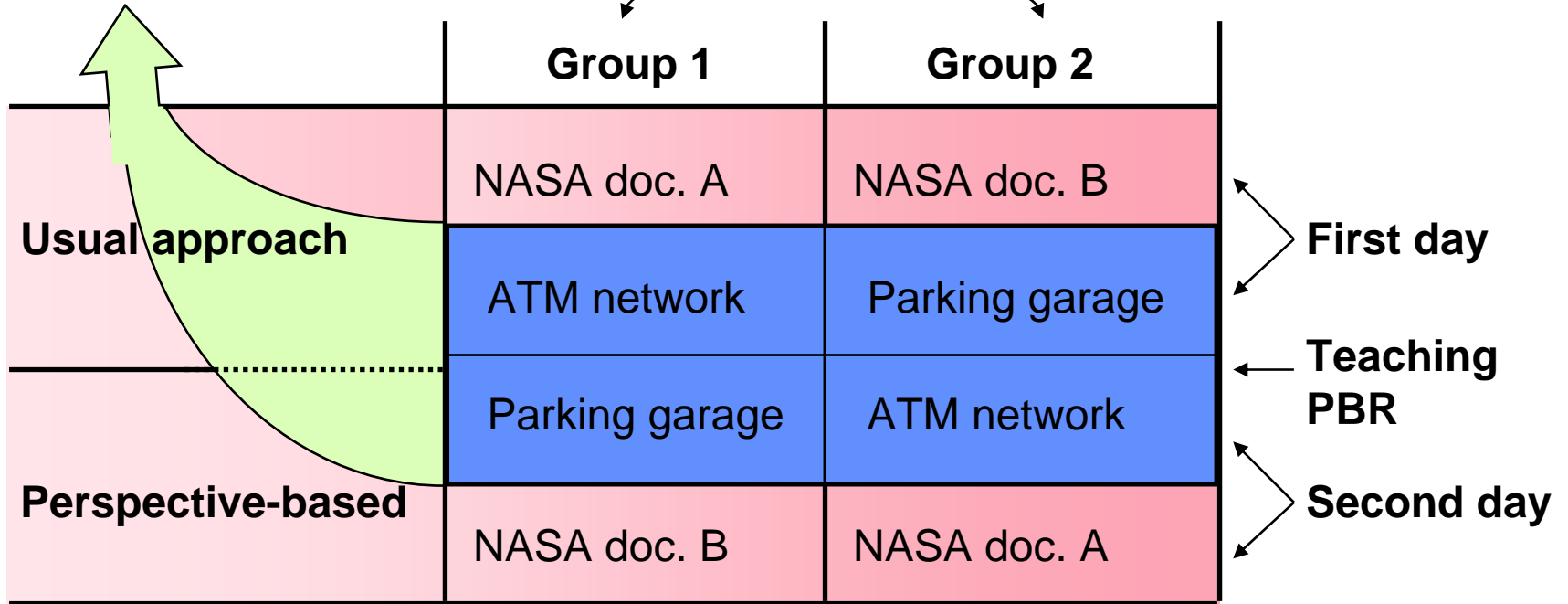
Subjects: 25 subjects in total



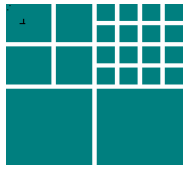
# Design of the PBR Experiment



Generic part



**Perspectives randomly and evenly assigned  
Training in front of every test**



# What is wrong with this design?

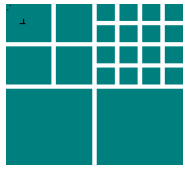
---

## **Internal issues?**

History, Maturation, Testing, Instrumentation, Statistical regression,  
Selection, Mortality, Selection-Maturation Interaction

## **External issues?**

Testing and X, Selection and X, Reactive Effects of Experimental  
Arrangement, Multi-treatment Interference



# Reading for Analysis: Blocked Subject-Project Study

---



## Defect-Based Reading

### Study Goal:

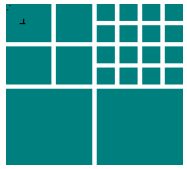
Analyze defect-based reading, ad-hoc reading and check-list based reading to evaluate and compare them with respect to their effect on fault detection effectiveness in the context of an inspection team from the viewpoint of quality assurance

### Environment:

University of Maryland graduate courses  
Requirements documents written in SCR notation

### Experimental design:

Blocked subject-project: Partial factorial design  
Replicated twice  
Subjects: 48 subjects in total  
Note: 1A – first replication, team A

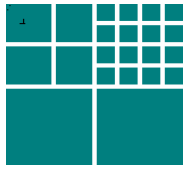


# Reading for Analysis: Blocked Subject-Project Study



## Defect-Based Reading

A	$O_1$		$O_2$		$O_3$	$X_2$	$O_4$
B	$O_1$	$X_1$	$O_2$		$O_3$	$X_0$	$O_4$
C	$O_1$	$X_2$	$O_2$		$O_3$		$O_4$
D	$O_1$		$O_2$		$O_3$	$X_1$	$O_4$
E	$O_1$	$X_1$	$O_2$		$O_3$	$X_2$	$O_4$
F	$O_1$	$X_2$	$O_2$		$O_3$		$O_4$
G	$O_1$	$X_1$	$O_2$	$X_2$	$O_3$	$X_1$	$O_4$
H	$O_1$	$X_2$	$O_2$	$X_1$	$O_3$	$X_2$	$O_4$



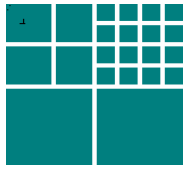
# Defect-Based Reading



## Round/Specification

**Detection  
Method**

	Round 1		Round 2	
	WLMS	CRUISE	WLMS	CRUISE
ad hoc	1B, 1D, 1G, 1H, 2A	1A, 1C, 1E, 1F, 2D	1A	1D, 2B
checklist	2B	2E, 2G	1E, 2D, 2G	1B, 1H
scenarios	2C, 2F	2H	1F, 1C, 2E, 2H	1G, 2A, 2C, 2F



# What is wrong with this design?

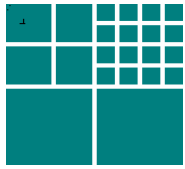
---

## **Internal issues?**

History, Maturation, Testing, Instrumentation, Statistical regression,  
Selection, Mortality, Selection-Maturation Interaction

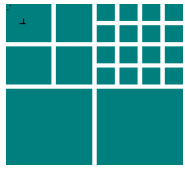
## **External issues?**

Testing and X, Selection and X, Reactive Effects of Experimental  
Arrangement, Multi-treatment Interference



# References

- V. Basili, Evolving and Packaging Reading Technologies, *The Journal of Systems and Software*, Volume 38, Number 1, pp. 312, July 1997.
- F. Shull, F. Lanubile, and V. Basili, Investigating Reading Techniques for Object-Oriented Framework Learning, *IEEE Transactions on Software Engineering*, Vol. 26, No. 11, pp. 1101-1118, November 2000.
- V. Basili, F. Shull, and F. Lanubile, Building Knowledge through Families of Experiments, *IEEE Transactions on Software Engineering*, Volume 25, Number 4, pp. 456-473, July 1999.
- Z. Zhang, V. Basili, and B. Shneiderman, Perspective-based Usability Inspection: An Empirical Validation of Efficacy, *Empirical Software Engineering: An International Journal*, Volume 4, No. 1, March 1999.
- V. Basili, S. Green, O. Laitenberger, F. Shull, S. Sorumgaard, and M. Zelkowitz, The Empirical Investigation of Perspective-based Reading, *Empirical Software Engineering, An International Journal*, Volume 1, Number 2, pp. 133-164, Kluwer Academic Publishers, October 1996.
- A.A. Porter, L.G. Votta, and V. Basili, Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment, *IEEE Transactions on Software Engineering*, Volume 21, Number 6, pp. 563-575, June 1995.



# References



- V. Basili, G. Caldiera, F. Lanubile, and F. Shull, "Investigating Focused Techniques for Understanding Frameworks", 1st International Workshop on Empirical Studies of Software Maintenance (WESS '96), pp. 49-53, Monterey, California, USA, 1996.
- F. Lanubile, F. Shull, and V. Basili, Experimenting with Error Abstraction in Requirements Documents, Presented at the Fifth International Symposium on Software Metrics, Bethesda, MD, November 18-20, 1998.
- V. Basili, F. Lanubile, and F. Shull, Investigating Maintenance Processes in a Framework-Based Environment, Presented at the International Conference on Software Maintenance, ICSM'98, Bethesda, MD, November 18-20, 1998.
- G. Travassos, F. Shull, M. Fredericks, and V. Basili, Detecting Defects in Object Oriented Designs: Using Reading Techniques to Increase Software Quality, ACM Sigplan Notices, Volume 34, Number 10, pp. 47-56, October 1999.
- F. Shull, I. Rus, and V. Basili, How Perspective-Based Reading Can Improve Requirements Inspections, IEEE Computer, Vol. 33, No. 7, July 2000.
- V. Basili, F. Shull, and F. Lanubile, Using Experiments to Build a Body of Knowledge, Proceedings of the Third International PSI Conference, Novosibirsk, Russia, pp. 265-282, July 1999.