Software Industry Experiments: A Systematic Literature Review

N. Juristo
O. Dieste
M. Danilo Martinez

Agenda

- Motivation
- Related work
- Goals
- Search strategy & results
- Results
- Interpretation
- Threats to validity

Motivation

- Nowadays, SE experiments are quite common in academia
- Unfortunately, lab experiments have strong limitations regarding generalizability (external validity)
  - Type of subjects, experience
  - Task accomplished
  - Settings

Field experiments are the natural complement to lab experiments

Motivation

- No survey exists about SE field experiments
  - We have no figures of industrial experiments
- We do not know
  - How many field experiments have been run
  - What factors they tested
  - Response variables, etc.

We wanted evidence of the intuitions we all have about experiments in software industry

Related work

- The only rigorous information on controlled experiments
  - *Sjøberg et al.* [1]

- Picture from *Sjøberg et al*
  - 26% (27/103) were done with professional participants
    - 17 do not report the type of environment (lab/field)
    - 7 have been run in the lab
    - 1 have been run in an industrial context

  - Experiments with professionals tend to have fewer number of subjects and less workload
    - Probably to put down the cost factor

Goals

- Figures
  - How many experiments have been run in industry?
  - what is the observed experiment time distribution?
- Experimental information
  - What independent variables do they study?
  - What dependent variables do they study?
  - What types of design do they use?
- Subjects
  - How many participate in industry experiments?
  - What categories of subjects participate in industry experiments?
- Lessons learnt
  - What challenges does experiments in industry raise?

Review planning and execution

- Systematic literature review (/scoping study)
  - Protocol development
  - Review execution and protocol refinement
  - Analysis of the gathered information
  - Reporting
- Analysis consisted in the tabulation and interpretation of the information acquired in the primary studies

Search

Strategy
Results

Search Strategy

- PICOC search string
  - We use Dieste, Griman & Juristo's recommendation for locating experiments [9]
- SCOPUS database
- Documents in English
- Published up until July 2012

Search substring (keywords linked by OR)	Keywords	PICOC term
1	Software	Population
2	Experiment Empirical Empirical study Empirical evaluation Experimentation Experimental comparison Experimental analysis Experimental evidence Experimental setting Empirical data	Intervention
3	Industry / Industries	Context
4	Company / Companies	
5	Business / Businesses	
6	Enterprise / Enterprises	

Search Results

- 658 recovered papers (including duplicates)
- 23 papers where pre-selected after applying the inclusion and exclusion criteria
- 15 papers remained after reading the full text

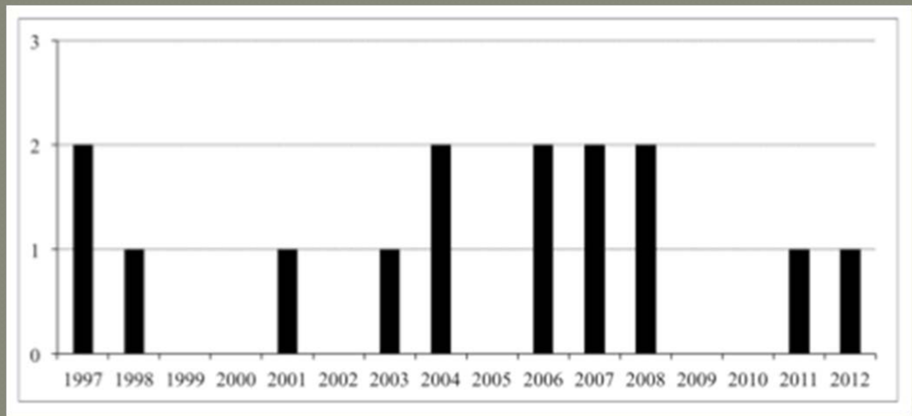
	Search strings[a]	Papers		
		Identified	*Pre-selected*	*Selected*
Planned	1 AND 2 AND 3	117	16	6
	1 AND 2 AND 4	129	11	4
	1 AND 2 AND 5	188	1	0
	1 AND 2 AND 6	64	4	0
Unplanned	1 AND 2 AND "Industrial"	160	16	8
	Total		39	15

Review results

Figures
Experimental Information
Subjects
Lessons Learnt

Figures

- We found 15 experiments run in industry
- Time distribution
  - 3 run before 2000
  - Most of them run between 2003-2008
  - We have taken back to pre-2000 levels



Experimental information

- Independent variables
  - 47% Quality
    - Inspection
    - Testing
  - 25% Management
    - Estimation
    - Agile

Area	*Technologies*	*Studies*	*Total*
Quality	Inspection	E3, E4, E7, E10, QE4	5
	Testing	E2, E6	2
Management	Estimation	QE5, QE3	2
	Agile	E1, E5	2
Object-Orientation	Object-oriented development	E8	1
	UML models	E9	1
	Others	QE1, QE2	2

Experimental information

- Independent variables
  - Over 50% of studies use widespread techniques
  - Techniques are not tested in industry for a good many years after they are invented
    - TDD and Pair are an exception

Technologies	*Techniques*	*Studies*	*Year of publication*
Inspection	Perspective-based reading	E4	1997
	Perspective-based reading Checklist-based reading	QE4	2001
	-	E3, E7, E10	1997, 2003, 2008
Testing	Test-driven development	E2, E6	2004, 2006
Estimation	-	QE5, QE3	1998, 2012
Agile	Pair design	E1	2007
	Pair programming	E5	2007
Object-oriented development	UML	E8	2004
UML models	UML	E9	2006
Others	-	QE1, QE2	2008, 2011

a. Years are given in increasing order and do not correspond to the column headed Studies

Experimental information

- Dependent variables
  - Three main response variables
    - Effectiveness (60% of studies)
    - Effort (33%)
    - Quality (27%)
  - These three variables refer to key business aspects
    - Their majority use is by no means surprising

Response variable	***Most common metrics***	***Studies***	***Total***
Effectiveness	Number of defects (9 cases)	E3, E4, E7, E8, E9, E10, QE1, QE4, QE5	9
Effort	Time (5 cases)	E1, E3, E5, E7, QE2	5
Quality	-	E1, E2, E5, E6	4
Others	-	E2, E6, QE3	3

Experimental information

- Type of Experimental designs
  - The most used is the full factorial (60%)
    - Given the low sample sizes, power is a concern
    - Risk of non-significant results
  - Low use of cross-over designs
    - very used in lab for increasing sample size
    - But increase also workload

Design		***Studies***	***Total***
Factorial	Full	E2, E3, E4, E5, E8, E9, E10, E7, QE1	9
	Fractional	QE2	1
Cross-over	Counterbalanced	E1, E6	2
	Unbalanced	QE4	1
Correlational study		QE3, QE5	2

Subjects

- Number of subjects
  - Experiments in industry have 69 subjects in average
    - Higher than the 20 subjects find by Sjøberg et al.'s
  - Some experiments have a very large sample size and bias the calculations
    - 26.8 is probably more accurate calculation

Subject type	*Studies*	*Total number of subjects*	*Average*
Professionals	E5, E8, E10, QE1	382	96.5
Software developers	E4, QE3, QE4	445	148.3
Engineers/software engineers	E1, QE2	26	13
Developers	E3, E7	21	11.5
Programmers	E2	24	-
Practitioners	E9	44	-
Employees	E6	28	-
Others	QE5	68	-
Total		1,038	69.2

Subjects

- Categories of subjects
  - We, like Sjøberg et al., find that names used to refer to professionals are very vague

Subject type	*Studies*	*Total number of subjects*	*Average*
Professionals	E5, E8, E10, QE1	382	96.5
Software developers	E4, QE3, QE4	445	148.3
Engineers/software engineers	E1, QE2	26	13
Developers	E3, E7	21	11.5
Programmers	E2	24	-
Practitioners	E9	44	-
Employees	E6	28	-
Others	QE5	68	-
Total		1,038	69.2

Lessons Learnt

- Challenges
  - Time
    - *The experiment might be assumed as time-consuming for the project, causing delay and hence being rejected*
  - Cost
    - *In many organizations it is hard to motivate experiments because organizations are concerned about financial issues*
  - Workload and planning
    - *The industrial reality at […] is very hectic, and pre-planning of all details was not feasible*
  - Academia vs. industry reality
    - *We realized that the term 'experiment' itself was demotivating because it focuses much more on the academic than the industrial benefit*

Interpretation & Threats

Results Interpretation

- The window of opportunity for running experiments in industry is very narrow and linked to four factors
  - Interference of experiments in production processes
    - Experiments should not be presented or allowed to be conceptualized as extra work
  - Alignment with business goals
    - The experiment should be run on a topic that is directly useful to the company
  - Human resource optimization
    - Experiments should take up as little of professionals' time as possible
  - Schedule flexibility
    - Experiments cannot be planned to a strict schedule, and execution times have to be flexible

Validity Threats

- Usage of SCOPUS database only
  - SCOPUS indexes publications from other databases like IEEE, ACM, Springer and Elsevier
    - Coverage is wide
  - Few studies conducted in industry are likely to be published in low-ranking media
    - This maximizes the likelihood of their being located in SCOPUS
- Search string
  - We have used pre-packaged search strings (ref. [9]) tailored for experiments in academia
  - But the terms for referring to experiments in industry can be many

Software Industry Experiments: A Systematic Literature Review

N. Juristo

Thank you !