Experimental Evaluation in Computer Science: A Quantitative Study

Paul Lukowicz, Ernst A. Heinz, Lutz Prechelt and Walter F. Tichy

Journal of Systems and Software
January 1995

Outline

- Motivation
- Related Work
- Methodology
- Observations
- Accuracy
- Conclusions
- Future work!

Introduction

- Large part of CS research new designs
  - systems, algorithms, models
- Objective study needs experiments
- Hypothesis
  - *Experimental study often neglected in CS*
- If accepted, CS inferior to natural sciences, engineering and applied math
- Paper ‘scientifically’ tests hypothesis

Related Work

- 1979 surveys say experiments lacking
  - 1994 say experimental CS under funded
- 1980, Denning defines experimental CS
  - *“Measuring an apparatus in order to test a hypothesis”*
  - *“If we do not live up to traditional science standards, no one will take us seriously”*
- Articles on role of experiments in various CS disciplines
- 1990 experimental CS seen as growing, but 1994
  - *“Falls short of science on all levels”*
- No systematic attempt to assess research

Methodology

- Select Papers
- Classify
- Results
- Analysis
- Dissemination (this paper)

Select CS Papers

- Sample broad set of CS publications (200 papers)
  - ACM Transactions on Computer Systems (TOCS), volumes 9-11
  - ACM Transactions on Programming Languages and Systems (TOPLAS), volumes 14-15
  - IEEE Transactions on Software Engineering (TSE), volume 19
  - Proceedings of 1993 Conference on Programming Language Design and Implementation
- Random Sample (50 papers)
  - 74 titles by ACM via INSPEC (24 discarded)
    - 30 refereed

Select Comparison Papers

- Neural Computing (72 papers)
  - Neural Computation, volume 5
  - Interdsciplinary: bio, CS, math, medicine …
  - Neural networks, neural modeling …
  - Young field (1990) and CS overlap
- Optical Engineering (75 papers)
  - Optical Engineering, volume 33, no 1 and 3
  - Applied optics, opto-mech, image proc.
  - Contributors from: ee, astronomy, optics…
  - Applied, like CS, but longer history

Classify

	Ernst	Lutz	Paul	Walter
NC		X		
OE			X	
TOCS			X	
Random	X	X	X	X
PLDI	X		X	
TOPLAS	X		X	
TSE			X	X

- Same person read most
- Two read all, save NC

Major Categories

- Formal Theory
  - Formally tractable: theorem's and proofs
- Design and Modeling
  - Systems, techniques, models
  - Cannot be formally proven → require experiments
- Empirical Work
  - Analyze performance of known objects
- Hypothesis Testing
  - Describe hypotheses and test
- Other
  - Ex: surveys

Subclasses of Design and Modeling

- Amount of physical space for experiments
  - Setups, Results, Analysis
- 0-10%, 11-20%, 21-50%, 51%+
- To shallow? Assumptions:
  - Amount of space proportional to importance by authors and reviewers
  - Amount of space correlated to importance to research
- Also, concerned with those that had no experimental evaluation at all

Assessing Experimental Evaluation

- Look for execution of apparatus, techniques or methods, models validated
- Tables, graphs, section headings…
- No assessment of quality
- But count only 'true' experimental work
  - Repeatable
  - Objective (ex: benchmark)
- No demonstrations, no examples
- Some simulations
  - Supplies data for other experiments
  - Trace driven

Outline

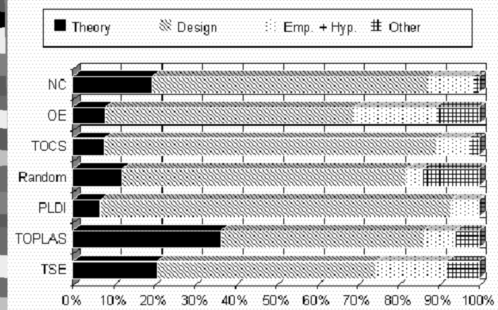
- Motivation
- Related Work
- Methodology
- *Observations*
- Accuracy
- Conclusions
- Future work!

Observation of Major Categories

	NC	OE	TOCS	Random	PLDI	TOPLAS	TSE
Theory	14	6	3	6	2	19	18
Design	40	46	31	35	23	26	47
Empirical	3	12	3	1	2	4	15
Hypothesis	0	3	0	1	0	0	0
Emp. + Hyp.	3	15	3	2	2	4	15
Other	1	8	1	7	0	3	7
Total	72	75	38	50	26	52	87

- Majority is design and modeling
- The CS samples have lower percentage of empirical work than OE and NC
- Hypothesis testing is rare (4 articles out of 403!)

Observation of Major Categories



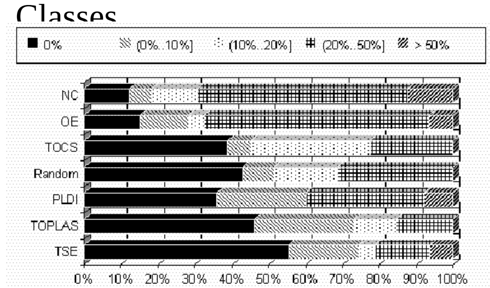
- Combine hypothesis testing with empirical

Observation of Design Sub-Classes

	NC	OE	TOCS	Random	PLDI	TOPLAS	TSE
0%	6	7	12	15	9	12	26
[0%..10%]	3	6	2	3	6	7	9
[10%..20%]	6	2	10	6	0	3	2
[20%..50%]	28	28	7	11	8	4	7
> 50%	6	3	0	0	2	0	3
Total	49	46	31	35	25	26	47
>20% / Total	69%	67%	23%	31%	40%	15%	21%
0% / Total	12%	15%	39%	43%	36%	46%	55%

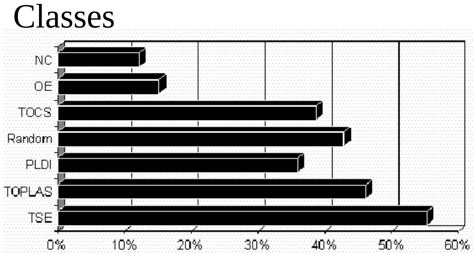
- Higher percentage with no evaluation for CS vs. NC+OE (43% vs. 14%)

Observation of Design Sub-Classes



- Many more NC+OE with 20%+ than in CS
- Software engineering (TSE and TOPLAS) *worse* than random

Observation of Design Sub-Classes



- Shows percentage that have 20%+ or more to experimental evaluation

Groupwork: How Experimental is WPI CS?

- Take 2 papers: PEDS, SERG, DSRG, ADVIS, REFER, AIRG
- Read abstract, flip through
- Categorize:
  - Formal Theory
  - Design and Modelling
    - + Count pages for experiments
  - Empirical
  - Hypothesis Testing
  - Other
- Swap with another group

Outline

- Motivation
- Related Work
- Methodology
- Observations
- *Accuracy*
- Conclusions
- Future work!

Accuracy of Study

- Deals with humans, so subjective
- Psychology techniques to get objective measure
  - Large number of users
    - → Beyond resources (and a lot of work!)
  - Provide papers, so other can provide data
- Systematic errors
  - Classification errors
  - Paper selection bias
- Statistical error

Systematic Error: Classification

	Theory	Empirical	Hypothesis	Other	Design				
					0%	(0%,10%]	(10%,20%]	(20%,50%]	>50%
Theory		0	3	2	24	2	6	1	0
Empirical	0%		1	0	2	1	2	4	0
Hypothesis	0%	1%		0	0	3	2	1	0
Other	2%	0%	0%		9	1	2	1	1
0%	26%	2%	0%	10%		6	4	9	0
(0%,10%]	2%	1%	0%	1%	6%		5	4	1
(10%,20%]	0%	2%	2%	2%	4%	5%		7	0
(20%,50%]	1%	4%	1%	1%	10%	4%	8%		1
>50%	0%	0%	0%	1%	0%	1%	1%	1%	

- Classification differences between 468 article classification pairs

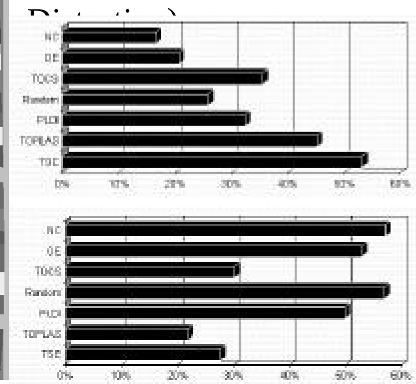
Systematic Error: Classification

- Classification ambiguity
  - Large between Theory and Design-0% (26%)
  - Design-0% and Other (10%)
  - Design-0% with simulations (20%)
- Counting inaccuracy
  - 15% from counting experiment space differently

Systematic Error: Paper Selection

- Journals may not be representative of CS
  - PLDI proceedings is a 'case study' of conferences
- Random sample may not be "random"
  - Influenced by INSPEC database holdings
  - Further influenced by library holdings
- Statistical error if selection within journals do not represent journals

Overall Accuracy (Maximize



No Experimental Evaluation

20%+ Space for Experiments

Conclusion

- 40% of CS design articles lack experiments
  - Non-CS around 10%
- 70% of CS have less than 20% space
  - NC and OE around 40%
- CS conferences no worse than journals!
- Youth of CS is not to blame
- Experiment difficulty not to blame
  - Harder in physics
  - Psychology methods can help
- Field as a whole neglects importance

Guidelines

- Higher standards for design papers
- Recognize empirical as first class science
- Need more publicly available benchmarks
- Need rules for how to conduct repeatable experiments
- Tenure committees and funding orgs need to recognize work involved in experimental CS
- Look in the mirror

Future Work

- Experiment in 1994 … how is CS today?
- 30 people in class
- 200 articles
- Each categorized by 2 people
- About 15 articles each

→Publish the results!

- (Send me email if interested)