The role of experimentation in computer science

Marvin Zelkowitz

Notes based upon a talk given in 2002-2004

So what is science?

How does science move from conjectures to established theories?

Essence of science

"All we can ask of a theory is to predict the results of events that can be measured. This sounds like an obvious point, but forgetting it leads to the so-called paradoxes that popular writers of our culture are fond of exploiting."

– Leon Lederman, Nobel Laureate physicist

Scientific theory

- a set of rules that relates quantities to observations we make
- an idea, model, or explanation that has been *tested* and accepted by the scientific community
- A good theory is characterized by making *predictions* that can be disproved or falsified by observations

Measurement crucial to science

"I often say that when you can measure what you are speaking about, and express it in numbers, you can know something about it. But when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind."

– Lord Kelvin

But we need relevancy

"The government is very keen on amassing statistics – they collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But what you must never forget is that every one of those figures comes in the first instance from the village watchman, who just puts down what he damn pleases."

– British economist Josiah Stamp, 1929

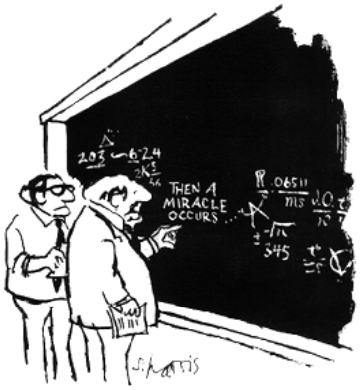
How science works …

"A rapid reciprocation of guesswork and checkwork, proposal and disposal, conjecture and refutation." – Sir Peter Medawar

"Flashes of inspiration are followed by rigorous test." – Thomas Gilovich

"You need a lot of ideas, and then you have to throw away the bad ones." – Linus Pauling – 2 time Nobel prize winner

The language of science - Mathematics



"I think you should be more explicit here in step two."

Scientific truth?

- Nothing said yet about "truth" in science. Science doesn't deal with truth.
- Science deals with obtaining the best predictions possible from observed data
- Future developments may change the underlying model as long as the observed relationships are maintained

So where does computer science come in?

- Computer science needs to operate within this scientific model of theory formation and experimental validation
- Software engineering's main laboratory is industrial developments
  - Investigating new technologies means to work with developers using new technologies
  - Goal is to transfer new technologies to industry

Speed of tech transfer influenced by

- The nature of the communication channels used to increase awareness and knowledge of the technology
- The nature of the social system in which the potential user operates
- The extent of efforts to diffuse the technology throughout an organization
- The technology's attributes
  - relative advantage
  - compatibility
  - complexity
  - trialability
  - observability

Delphi technique

- A group of experts receives the specification plus an estimation form.
- The experts discuss product and estimation issues.
- The experts produce individual estimates.
- The estimates are tabulated and returned to the experts.
- An expert is made aware only of his or her own estimate; the sources of the remaining estimates remain anonymous.
- The experts meet to discuss the results.
- The estimates are revised.
- The experts cycle through steps 1 to 7 until an acceptable degree of convergence is obtained.

Delphi technique -2

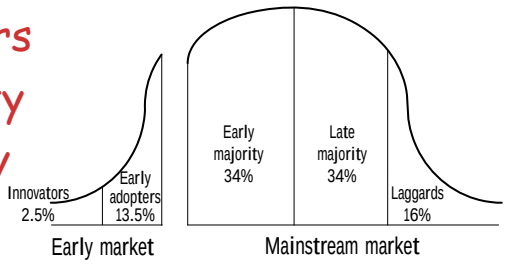
Perceived strengths and weaknesses of the Delphi Technique

Weaknesses	Strengths
can wrongly influence an individual and the impact of a dominant individual	experts with different backgrounds/perspectives
depends upon knowledge/expertise of individuals	group discussion can correct mistakes
risk of erroneous assumptions	reconsideration
group discussion made little difference to the result (consensus group)	uses expert judgment
high variability in predictions	median better than mean
inappropriate target, should use for more detailed problems	provides comparison with other estimates
	anonymity/independence combined with group benefits

Technology transfer

Players:

- Innovators
- Early adopters
- Early majority
- Late majority
- Laggards

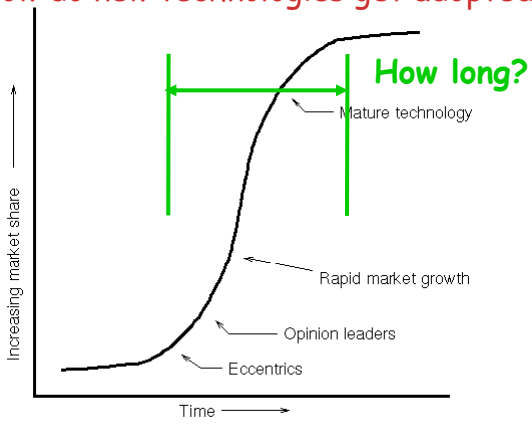


Types of evidence

- Tangible evidence
- Testimonial evidence
- Equivocal testimonial evidence
- Missing evidence
- Accepted facts

Technology transfer model S-curve growth

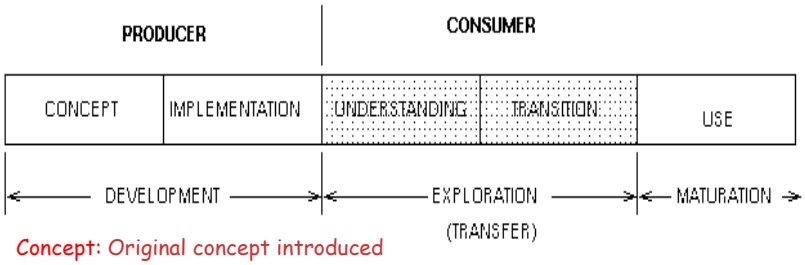
How do new technologies get adopted?:



Redwine-Riddle study

- Technology maturation takes time:
- From Redwine – Riddle study (1985):
- Studied 17 software engineering technologies of the 1960s and 1970s (e.g., spreadsheets, UNIX)
- Required an average of 17 years from concept to maturation
- Required an average of 7.5 years after initial development to widespread availability in industry
- Similar times compared to other engineering disciplines

Technology maturation life cycle (from Redwine and Riddle)



Concept: Original concept introduced
Implementation: Initial implementation of technology
Exploration (understanding): Others experiment with technology, expand and modify it
Exploration (transition): Technology spreads across industry
Use: Mature when 70% of industry uses it
Technologies generally require 17-25 years to mature.
Corporate infusion of a new technology generally required 5-7.5 years

But how does new technology get validated?

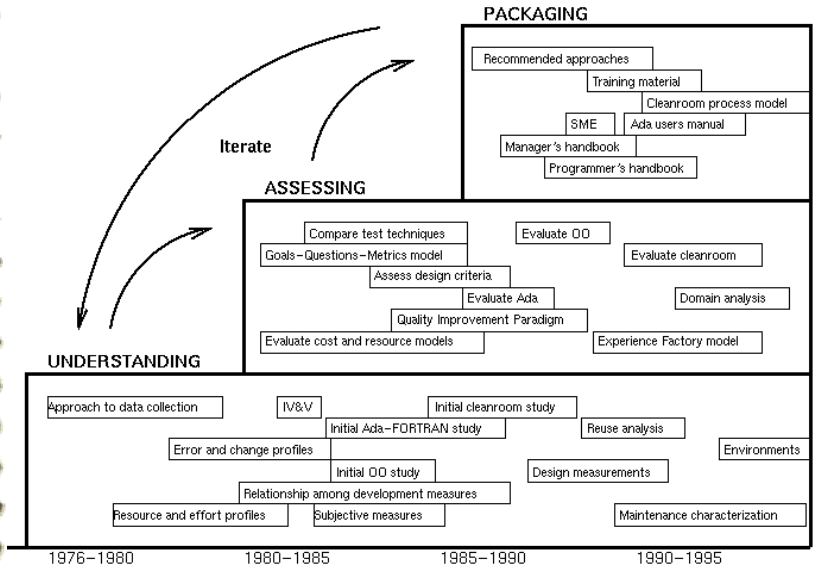
- ✓ Lots of technology development
- ✓ Rapid change today within our technological society
- ✓ But software failures are all too common
- ✓ Why such failures?
- ✓ We need research laboratories for software engineering
  - NASA Software Engineering Laboratory (1976-2002) one such example

Technology transfer experience from the NASA Software Engineering Laboratory

Background of NASA/GSFC SEL:

- Began in 1976 to study software development
- Typical applications: ground support software for unmanned spacecraft
- Characteristics:
  - Size from 10K to 500K source lines
    - 1970-1990 :FORTRAN dominant language; 1990-2000; C and C++
  - Typically 10-15 people for 18--24 months
  - Mixture of contractor and government personnel
  - Over 125 projects; 500MB Oracle database
- Many studies of effects of process changes on development in SEL environment

NASA technology transfer process



Software engineering technology transfer

Technology transfer is generally product oriented:

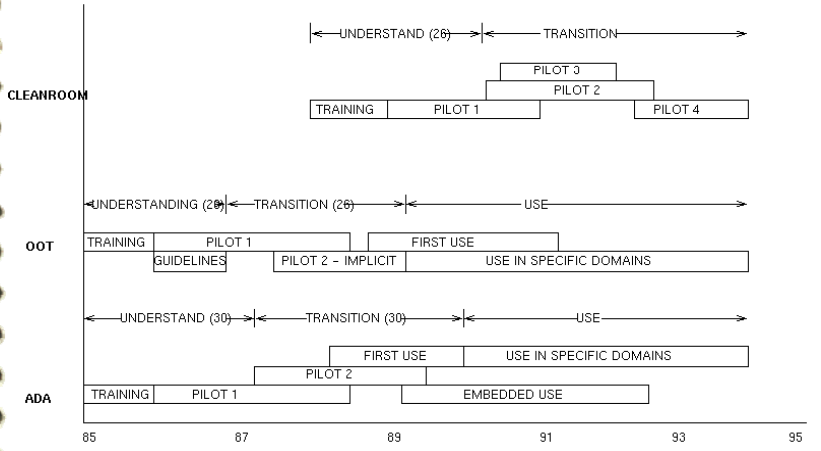
- In most engineering disciplines, the process is centered in a product.

Software engineering does not yet achieve that – Processes describing actions to take are as important as the tools that are used.

For example, many of the technologies explored by the SEL are procedures only and not tools:

- Object oriented technology
- Goals/Question/Metrics model
- Measurement
- Cleanroom
- Inspections

Examples of technology infusion



Examples of transferred technologies

Survey of software professionals – What 10 technologies (out of a list of over 100) have helped your productivity the most?

TOTAL REPLIES	44	FROM NASA	12
Workstations,pcs	27	Object oriented	12
Object oriented	21	Networks	10
GUIs	17	Workstations,pcs	8
Process models	16	Process models	7
Networks	16	Measurement	5
C and C++	8	GUIs	4
CASE tools	8	Structured design	3
Databases	8	Databases	2
Desktop publish	8	Desktop publish	2
Inspections	7	Development meth	2
Email	7	Reuse	2
Measurement	6	Cost estimation	2
		Comm. Software	2

What is the problem in validating a new technology?

- Often there is a lack of validation before using a new technology
  - Anecdotal evidence that we don't validate our claims
  - Study by Tichy (1995) that 50% of software engineering papers do not have validation;
  - Only 15% in other scientific fields
- Going back to the definitions of science, we need measurements

Experimental models for software research

- But in Computer Science:
  - Our theories are our tools and techniques
  - All too often, we don't appreciate the "science" in our title
  - Validation, experimentation, and measurement seem to be lacking
- Recognition that we need to understand how to experiment in software engineering
- Problems:
  - Models mostly taken from social science domain.
  - View experimentation as the replication of a hypothesis under varying controlled conditions

Can we take larger view of experimentation that applies in the software domain?

Experiment taxonomy

- Replicated experiments
  - Chemistry – Rows of test tubes
  - Psychology – Freshmen students working on a task
- Observations
  - Medicine – Clinical trials
  - Astronomy – Observe events if and when they occur
- Data Mining of completed activities
  - Archaeology – Dig up the past
  - Forensic investigations – recreate what happened
- How do these relate to Software?
- What data does each method generate?

Basic data collection models

Impact on the process being studied:

- Active methods – An effect on the process being studied
- Passive methods – No effect on process being studied

Based upon work of M. Zelkowitz and D. Wallace about 1995 at NIST

Classes of methods

- Controlled method – Multiple instances of an observation in order to provide for statistical validity of the results. (Usually an active method.)
- Observational method – Collect relevant data as it develops. In general, there is relatively little control over the development process. (Weakly active, although may be passive.)
- Historical method – Collect data from completed projects. (Passive methods.)

- These three basic methods have been classified into 12 data collection models.

(We will also consider one theoretical validation method, yielding 13 validation methods)

Controlled methods

Replicated – Several projects are observed as they develop (e.g., in industry) in order to determine the effects of the independent variable. Due to the high costs of such experiments, they are extremely rare.

Synthetic environments – These represent replicated experiments in an artificial setting, e.g., often in a university.

Dynamic analysis – The project is replicated using real project data.

Simulation – The project is replicated using artificial project data.

The first 2 of these generally apply to process experiments while the last two generally apply to product experiments.

Observational methods

Project monitoring – Collect data on a project with no preconceived notion of what is to be studied.

Case study – Data collected as a project develops by individuals who are part of the development group. (Often used in SEL.)

Field Study – An outside group collects data on a development. (A weaker form of case study.)

Historical methods

Literature search – Review previously published papers in order to arrive at a conclusion. (e.g., Meta-analysis - combining results from separate related studies)

Legacy data – Data from a completed project is studied in order to determine results.

Lessons-learned data – Interviews with project personnel and a study of project documentation from a completed project can be used to determine qualitative results. (A weak form of legacy data.)

Static analysis – Artifacts of a completed project are processed to determine characteristics.

But list of methods is incomplete

- Assertions: What do software engineers often do?
  - For a new technology validation often consists of:
    "I tried it and I like it"
  - Validation often consists of a few trivial examples of using the technology to show that it works.
  - Added this validation as a weak form of case study under the "Observational Method:"
- Assertion – A simple form of case study that does not meet rigorous scientific standards of experimentation.
- Theoretical validation – A form of validation based upon mathematical proof.

Summary of validation methods

Summary: 13 methods

- 11 experimental methods
- assertion (weak experimental validation)
- theoretical validation

Evaluation of this classification

Review of 1995 Tichy study:

- Reviewed 403 papers
- Sources: ACM journals and conferences, IEEE TSE
- Classification of papers
  - Formal theory - Proofs
  - Design and modeling- Designs which are not formal
  - Empirical study- Evaluation of existing technology
  - Hypothesis testing- Experiments to test a hypothesis
  - Other- Anything else, e.g., surveys

Conclusions from Tichy study

Those relevant to current study:

- 40% of computer science papers without validation
- 50% of software engineering papers without validation
- Comparable numbers are neuroscience (12%) and optical engineering (15%)
- But only considered design and modeling papers.

Perhaps too narrow a view of what is an experiment.

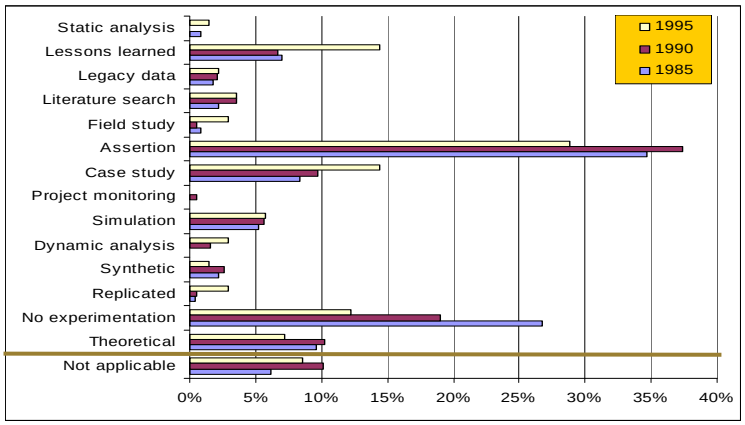
NIST evaluation

- Performed by Zelkowitz and Dolores Wallace
- New literature search: Papers from 1985, 1990, 1995
- Sources: 612 papers reviewed
  - IEEE Software --- a technical magazine
  - Transactions on Software Engineering - research journal
  - ICSE proceedings --- a conference
- Can we detect changing trends over 10 years?
- Added 2 more classifications to above 13:
  - Not applicable --- The paper does not discuss a new technology, e.g., a survey paper.
  - No experimentation --- The paper presents a new technology, but makes no claims as to experimental validity. These are the papers that SHOULD have validation of some form.

Summary of paper classifications

SUMMARY TOTALS	85			90			95			Ttl
Method	ICSE	Soft	TSE	ICSE	Soft	TSE	ICSE	Soft	TSE	
Not applicable	6	6	3	4	16	2	5	7	1	50
Theoretical	3	1	18	1	0	19	3	0	7	52
No experimentation	13	10	38	7	8	22	7	3	7	115
Replicated	1	0	0	0	0	1	1	0	3	6
Synthetic	3	1	1	0	1	4	0	0	2	12
Dynamic analysis	0	0	0	0	0	3	0	0	4	7
Simulation	2	0	10	0	0	11	1	1	6	31
Project monitoring	0	0	0	0	1	0	0	0	0	1
Case study	5	2	12	7	6	6	4	6	10	58
Assertion	12	13	54	12	19	42	4	14	22	192
Field study	1	0	1	0	0	1	1	1	2	7
Literature search	1	1	3	1	5	1	0	3	2	17
Legacy data	1	1	2	2	0	2	1	1	1	11
Lessons learned	7	5	4	1	4	8	5	7	8	49
Static analysis	1	0	1	0	0	0	0	0	2	4
Yearly totals	56	40	147	35	60	122	32	43	77	612

Classification of 612 papers



Quantitative observations

- Most prevalent validation mechanisms were lessons learned and case studies, each about 10%
- Simulation was used in about 5% of the papers, while the remaining techniques were each used in under 3% of the papers
- About one-fifth of the papers had no experimental validation
- Assertions (a weak form of validation) were about one-third of the papers
- But percentages of no experimentation dropped from 26.8% in 1985 to 19.0% in 1990 to only 12.2% in 1995. (Perhaps a favorable trend?)

Qualitative observations

- We were able to classify every paper according to our 13 categories, although somewhat subjective (e.g., assertion versus case study).
- Some papers can apply to 2 categories. We chose what we believed to be the major evaluation category.
- Authors often fail to clearly state what their paper is about. Its hard to classify the validation if one doesn't know what is being validated.
- Authors fail to state how they propose to validate their hypotheses.
- Terms (e.g., experiment, case study, controlled experiment, lessons learned) are used very informally.

Major caveat

The papers that appear in a publication are influenced by the editor of that publication or program committee. The editors and program committees from 1985, 1990, and 1995 were all different. This then imposes a confounding factor in our analysis process that may have affected our outcome.

Overall observations

- Many papers have no experimental validation at all (about one-fifth), but fortunately, this number seems to be dropping.
- BUT too many papers use an informal (assertion) form of validation. Better experimental design needs to be developed and used.
- Lessons learned and case studies each are used about 10% of the time, the other techniques are used only a few percent at most.
- Terminology of how one experiments is sloppy. We hope a classification model, such as this one, can help to encourage more precision in the describing of empirical research.

Comparison to other fields

- We decided to look at several other disciplines for comparison, An informal study. No attempt at choosing the "best" journal in each field.
- Journals:
  - J 1 – Measurement Science and Technology, (Devices to perform measurements)
  - J 2 – American Journal of Physics, (Theory and application of new physical theories)
  - J 3 – Journal of Research of NIST, (Research on measurement and standardization issues)
  - J 4 – Management Science, (Queueing theory and scheduling problems)
  - J 5 – Behavior Therapy, (Clinical therapies)
  - J 6 – Journal of Anthropological Research, (Study of human cultures)

Summary of paper classifications

Method	J1%	J2%	J3 %	J4 %	J5 %	J6%	TTL	%
NA		2	5			1	8	---
No exper+theory	16	58	7	21	6	31	26	20
Replicated		5	4	4	12		5	
Synthetic			4	11	29		9	
Dynamic anal.	32	5	19	11			17	
Simulation			15	32			13	
Proj. Mon.								
Case study	40	16	41		6	8	26	
Assertion	8	4	11			8	7	5
Field study				4	18		4	
Liter. Search	4	11	7	7	24	23	14	
Legacy data					6	23	4	
Lessons learn		5				8	2	
Static anal.								
Paper count(#)	25	21	32	28	17	14	137	

Note clustering of techniques across journals
No attempt to summarize across fields, except for experimentation and assertions

Results from other fields

- No experimentation plus assertion data much lower than in software engineering (25% versus 55%)
- Each field has a characteristic data collection model:
  - Physics --- dynamic analysis and simulation (repeated experiments
  - Psychology --- replicated and synthetic (repeated trials of individuals)
  - Anthropology --- legacy data (historical data)
- Literature search more accepted model for publication. (Does this refer to publication of similar studies that are frowned upon in computer science?)

NIST survey update

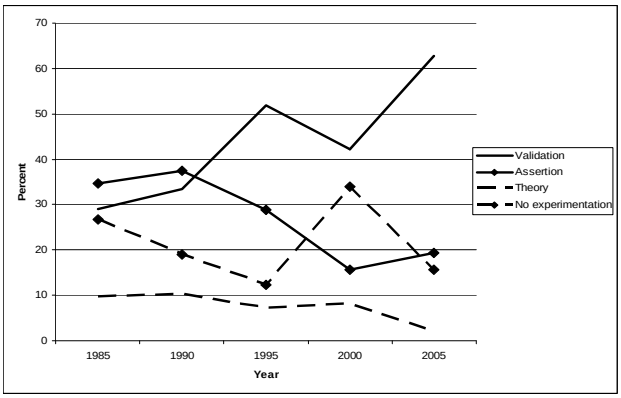
- Since study was done in 1995, two more data points since then: 2000 and 2005.
- Can we validate any of the claims of the earlier paper?
- Finally an update was studied in 2007

2007 data

Table 1. Basic classification data from 958 papers: 1985-2005.

		Project monitoring	Case study	Field study	Literature search	Legacy	Lessons learned	Static analysis	Replicated	Synthetic	Dynamic analysis	Simulation	Assertion	Theoretical	No experimentation	Not applicable	Total
	icse	0	5	1	1	1	7	1	1	3	0	2	12	3	13	6	56
	tse	0	12	1	3	2	4	1	0	1	0	10	54	18	38	3	147
	sw	0	2	0	1	1	5	0	0	1	0	0	13	1	10	6	40
1985 Total		**0**	**19**	**2**	**5**	**4**	**16**	**2**	**1**	**5**	**0**	**12**	**79**	**22**	**61**	**15**	**243**
	icse	0	7	0	1	2	1	0	0	0	0	0	12	1	7	4	35
	tse	0	6	1	1	2	8	0	1	4	3	11	42	19	22	2	122
	sw	1	6	0	5	0	4	0	0	1	0	0	19	0	8	16	60
1990 Total		**1**	**19**	**1**	**7**	**4**	**13**	**0**	**1**	**5**	**3**	**11**	**73**	**20**	**37**	**22**	**217**
	icse	0	4	1	0	1	5	0	1	0	0	1	4	3	7	5	32
	tse	0	10	2	2	1	8	2	3	2	4	6	22	7	7	1	77
	sw	0	6	1	3	1	7	0	0	0	0	1	14	0	3	7	43
1995 Total		**0**	**20**	**4**	**5**	**3**	**20**	**2**	**4**	**2**	**4**	**8**	**40**	**10**	**17**	**13**	**152**
	icse	0	10	0	0	1	4	0	2	2	4	1	11	3	20	10	68
	tse	0	9	3	1	0	0	0	0	4	4	7	11	10	15	2	66
	sw	0	6	4	0	0	0	0	0	1	3	1	3	0	19	14	51
2000 Total		**0**	**25**	**7**	**1**	**1**	**4**	**0**	**2**	**7**	**11**	**9**	**25**	**13**	**54**	**26**	**185**
	icse	0	14	1	0	1	0	0	0	3	8	1	10	1	3	0	42
	tse	0	9	4	1	5	0	2	1	2	13	5	13	1	8	2	66
	sw	0	7	3	1	1	3	0	0	3	0	0	4	1	12	19	53
2005 Total		**0**	**30**	**8**	**2**	**7**	**3**	**2**	**1**	**8**	**21**	**6**	**27**	**3**	**22**	**21**	**161**

Basic results



- Validation rose from 29% to 65%
- No experimentation dropped from 27% to 16%
- Assertions dropped from 35% to 19%

Trend continues to improve

Why doesn't industry "buy" this validation?

Industry:

- Ignores results from archival journals
- Believes in unsubstantiated rumors

Research community:

- Doesn't require validation
- Doesn't perform validations as thorough as necessary

There is a "disconnect" between these 2 cultures

Industrial methods

Method	Type	Method	Type
Case study	Observ	Literature search	Hist
Demonstrator projects	Contrl	Pilot study	Contrl
Education	Hist	Project monitoring	Observ
External	Informal	Replicated project	Contrl
Expert opinions	Hist	Synthetic benchmark	Contrl
Feature benchmark	Hist	Theoretical analysis	Formal
Field study	Observ	Vendor opinion	Informal
Legacy data	Hist		

Based on paper by Binkley, Wallace and Zelkowitz

Industrial methods-1

Additional methods often used by industry:

Expert opinion – use the opinion of experts. This can take the form of hired consultants brought in to teach a new technology or attendance of a trade show where various vendors demonstrate their products.

Edicts – changes required by an outside agent.

Feature analysis – a study of the features of the new technology and a subjective evaluation of its impact on the development process. Often used to compare two alternatives.

Industrial methods - 2

Compatibility studies – studies used to test whether various technologies can be combined or if they interfere with one another.

model problems – narrowly defined problems that the technology can address.

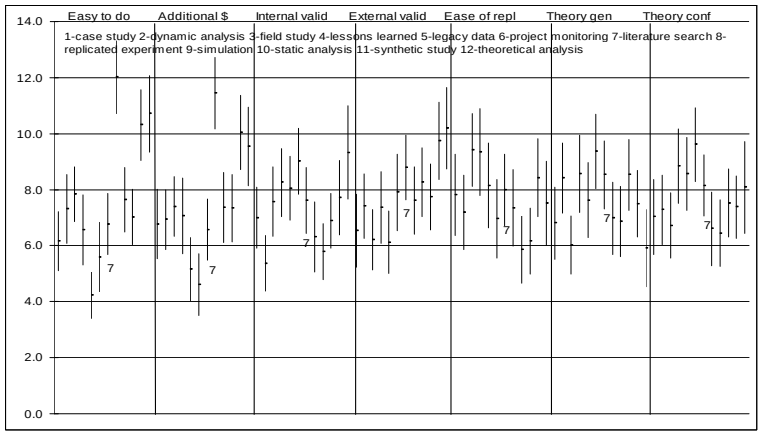
demonstrator study – scaled-up application development, with some attributes (e.g., performance, documentation) reduced in order to limit costs or development time.

Pilot study - This is a full-scale implementation using the new technology.

Relationship between methods

Research exploratory methods	Industrial confirmatory methods
Assertion	Vendor opinion
Case study	Case study
Dynamic analysis	Synthetic benchmarks
Field study	Field study
Legacy data	Legacy data
Lessons learned	Expert opinion
Literature search	Literature search
Project monitoring	Project monitoring
Replicated	Replicated project
Simulation	Pilot study
Static analysis	Feature benchmark
Synthetic	Demonstrator projects
Theoretical analysis	Theoretical analysis
None	Education
None	External

Survey on relative importance of each method (e.g., 18 from research community)



Survey results

Practical and impractical techniques from research sample

	Ease of use	Addit. $	Int. val.	Ext. val.	Ease of repl.	Theory gen.	Theory conf.
Practical	Dyn. Anal Les. Learned Legacy data Static anal.	Legacy data Proj. mon. Static anal.	Dyn. anal. Replicated		Dyn. anal. Simulation Static anal.		Replicated
Impractical	Replicated Synthetic	Replicated	Case study		Case study Field study Les. learned		Legacy data

Practical and impractical techniques from developer sample

	Ease of use	Addit. $	Int. val.	Ext. val.	Ease of repl.	Theory gen.	Theory conf.
Practical	Case study Legacy data Proj. mon.	Case study Legacy data Proj. mon. Lit. search	Case study Dyn. Anal. Simulation	Case study Legacy data	Case study	Case study Field study Theory anal.	Field study
Impractical	Replicated Synthetic Theory anal.	Replicated Synthetic Theory anal.	Proj. mon. Theory anal.	Synthetic Theory anal.		Proj. mon.	Proj. mon.

In conclusion …

- We have proposed a 13-way approach toward developing a quantitative model of software experimentation. It seems applicable to the software engineering literature.
- In a 1992 report from the National Research Council the Panel on Statistical Issues and Opportunities for Research in the Combination of Information recommended:

  "The panel urges that authors and journal editors attempt to raise the level of quantitative explicitness in the reporting of research findings, by publishing summaries of appropriate quantitative measures on which the research conclusions are based …''
- Researchers and practitioners have a different view of the world with respect to validating a technology
- In general, software engineering experimental validation is probably not as bad as folklore says, but could stand to do a better job.