

The Science of Software Engineering

Marvin Zelkowitz
Department of Computer Science
University of Maryland
College Park, Maryland
marv@zelkowitz.com

ESEM - 2009

ESEM - October 2009

1

A Need for a ~~The~~ Science of Software Engineering

Marvin Zelkowitz
Department of Computer Science
University of Maryland
College Park, Maryland
marv@zelkowitz.com

ESEM - 2009

ESEM - October 2009

2

Organization of talk

- Some personal comments on how I arrived at the theme of this talk
- What are the issues in developing a science of software engineering?
- What's next?

So what have I been doing for the past 40 years?

- Most of my professional life has been at the University of Maryland, teaching and doing research in the general area of software engineering.
- But those who know me, know that I have three other areas of great interest.

One is attending science fiction conventions



ESEM – October 2009

5

A second is my interest in model railroading



Layout obviously unfinished.

ESEM – October 2009

6

A third is that I consider myself a professional skeptic



ESEM – October 2009

7

A third is that I consider myself a professional skeptic

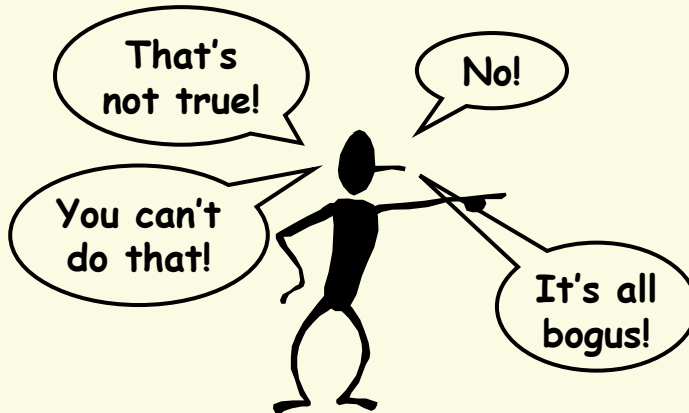
I belong to an organization of skeptics.



ESEM – October 2009

8

A third is that I consider myself a professional skeptic



A third is that I consider myself a professional skeptic

What does this really mean?

And how does this relate to software engineering?

This is the general theme of the rest of this talk.

Common view of skeptics - There's no way to help them



ESEM – October 2009

11

Are skeptics cynics?

- **Cynic** - One who shows a disposition to disbelieve in the sincerity or goodness of human motives and actions, and is wont to express this by sneers and sarcasms.
- **Skeptic** - one with doubt or incredulity as to the truth of some assertion or supposed fact.
Oxford English Dictionary

A cynic disbelieves everything, a skeptic wants to be convinced.

ESEM – October 2009

12

Why is skepticism important?

Conspiracy theories? Reincarnation?
Evolution? Global warming?
 Alternative medicine?
UFOs? Dowsing? Spirits and ghosts?
Moon Landing was a hoax?
Alien abduction? Natural remedies?
 Therapeutic touch?

Why is skepticism important?

Con... on?
Evo...
UF... osts?
Me...
Alien... es?

What do you believe?
What should your opinion be about these?

Emphasis on Critical Thinking

NCAS' mission: NCAS is an independent nonprofit educational and scientific organization that promotes critical thinking and scientific understanding, with a focus on paranormal and fringe-science claims. NCAS ... serves as an advocate for science and reason, actively promoting the scientific method, rational inquiry, and education.

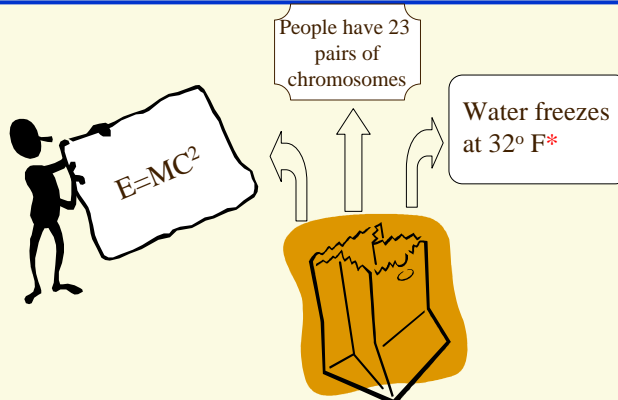
Emphasis on proper use of science and the scientific method in everyday life

"Science is a way of thinking, much more than it is a body of facts." - Carl Sagan

ESEM – October 2009

15

Does science tell us reality?



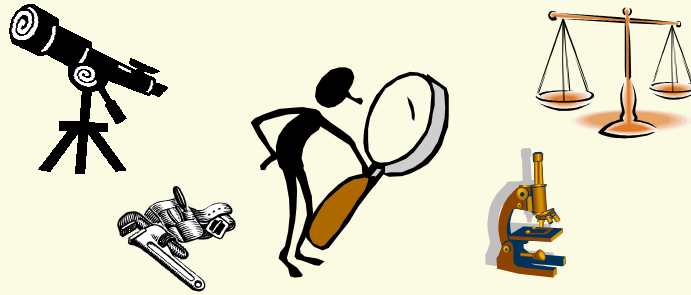
Science is not a bag of facts

*-Unless you are one of the 6 billion non-Americans, where water freezes at 0° C

ESEM – October 2009

16

Does science tell us reality?



~~Science is not a bag of facts~~ - it is a process to understand the world

ESEM - October 2009

17

Bad Science

- "Voodoo science" (Bob Park)
 - **Bad science** - Applying scientific method incorrectly
 - **Mistakes** - Performing an experiment incorrectly
 - **Fraud** - Intentional deception
- **Pseudoscience** - Violating accepted principles of the world (e.g., perpetual motion vs. thermodynamics)

ESEM - October 2009

18

Methods of pseudoscience

(see: <http://www.don-lindsay-archive.org/skeptic/arguments.html>)

Ad Hominem	Argument By Question
Affirming The Consequent	Argument By Repetition (Ad Nauseam)
Amazing Familiarity	Argument by Rhetorical Question
Ambiguous Assertion	Argument By Scenario
Appeal To Anonymous Authority	Argument By Selective Observation
Appeal To Authority	Argument By Selective Reading
Appeal To Coincidence	Argument By Slogan
Appeal To Complexity	Argument From Adverse Consequences
Appeal To False Authority	Argument From Age
Appeal To Force	Argument From Authority
Appeal To Pity	Argument From False Authority
Appeal To Widespread Belief	Argument From Small Numbers
Argument By Emotive Language	Argument From Spurious Similarity
Argument By Fast Talking	Argument Of The Beard
Argument By Generalization	Argument To The Future
Argument By Gibberish	Bad Analogy
Argument By Half Truth	Begging The Question
Argument By Laziness	Burden Of Proof
Argument By Personal Charm	Causal Reductionism
Argument By Pigheadedness	Changing The Subject
Argument By Poetic Language	Cliche Thinking
Argument By Prestigious Jargon	Common Sense

ESEM – October 2009

19

Pseudoscience-2

Complex Question (Lying)	Meaningless Questions
Confusing Correlation And Causation	Misunderstanding Statistics
Disproof By Fallacy	Moving The Goalposts
Equivocation	Needling
Error Of Fact	Non Sequitur
Euphemism	Not Invented Here
Exception That Proves The Rule	Outdated Information
Excluded Middle	Pious Fraud
Extended Analogy	Poisoning The Wells
Failure To State	Psychogenetic Fallacy
Fallacy Of Composition	Reductio Ad Absurdum
Fallacy Of Division	Reductive Fallacy (Oversimplification)
Fallacy Of The General Rule	Reifying
Fallacy Of The Crucial Experiment	Short Term Versus Long Term
False Cause	Slippery Slope Fallacy
False Compromise	Special Pleading (Stacking The Deck)
Genetic Fallacy	Statement Of Conversion
Having Your Cake	Stolen Concept
Hypothesis Contrary To Fact	Straw Man
Inconsistency	Two Wrongs Make A Right
Inflation Of Conflict	Weasel Wording
Internal Contradiction	
Least Plausible Hypothesis	
Lies	

ESEM – October 2009

20

Does science tell us reality?

- A scientific theory is characterized by making *predictions* that can be disproved or falsified by observations; Nothing is ever said about "truth" - Truth and falsity are philosophical concepts.
- Example: We don't know why or how gravity works
 - But we are quite sure if we step off the roof of a building, we will fall
 - And we are quite sure we know how long it will take and how fast we will hit the ground

Einstein's relativity revised Newton's theory of gravity with better predictions, but it still has to include explaining why stepping off a roof will make you go "splat" when you hit the ground.

How does this relate to software engineering? Lets first discuss - Homeopathy

- Conceived by Dr. Samuel Hahnemann in 1810
- Based on Law of similars - "Like cures like"
 - Any material that causes a reaction can be used, *if sufficiently dilute*, to eliminate that reaction
 - If pepper makes you sneeze, then a sufficiently dilute solution of pepper can cure the sneezing from allergies

By "dilute" we mean really really really dilute

- 1X dilution - 1 in 10, 2X dilution - 1 in 100 = 1:10² ...
20X dilution - 1:10²⁰
 - But only 6.023x10²³ molecules per mole (e.g., for water it would be 18 grams)
 - So at dilutions of 30X, only 1 chance out of about 1,000,000 that even 1 molecule of substance is present in one cup of solution
 - Most homeopathic solutions are at least 1 C = 100X = 1:10¹⁰⁰
- Homeopathy is big business today. Billions of dollars annually in the USA; used worldwide


Identifying Pseudoscience -

Use of vague, exaggerated or untestable claims		
Over reliance on confirmation rather than refutation		
Lack of openness to testing by other experts		
Absence of progress		
Personalization of issues; Proof by authority		
Lack of scientific method		

Identifying Pseudoscience - Homeopathy		
Use of vague, exaggerated or untestable claims	Dilutions of 1C non-measurable; The higher the dilution, the better the effect; No rational underlying theory; Only proponents can "see" effect	
Over reliance on confirmation rather than refutation	Testimonials on effectiveness. No blind studies of effects	
Lack of openness to testing by other experts	Allowed to be sold in USA. FDA prevented by law from studying its effectiveness	
Absence of progress	No change in "theory" since 1810	
Personalization of issues; Proof by authority	Homeopathy = Hahnemann	
Lack of scientific method	No multiple controlled studies	
ESEM – October 2009		25

Identifying Pseudoscience - Homeopathy		
Use of vague, exaggerated or untestable claims	Dilutions of 1C non-measurable; The higher the	
Over conf refu	<p>I realized about 15 years ago, that my interest in skepticism and working on experimental software engineering were really the same thing.</p> <p>How has software engineering changed over the past 20 years?</p> <p>What needs to be done to further improve the field?</p>	
Lack by o		
Abs		
Pers		
Proof by authority		
Lack of scientific method	No multiple controlled studies	
ESEM – October 2009		26

Identifying Pseudoscience - Software Engineering?

Use of vague, exaggerated or untestable claims	Dilutions of IC non-measurable; The higher the dilution, the better the effect; No rational underlying theory; Only proponents can "see" effect	
Over reliance on confirmation rather than refutation	Testimonials on effectiveness. No blind studies of effects	
Lack of openness to testing by other experts	Allowed to be sold in USA. FDA prevented by law from studying its effectiveness	
Absence of progress	No change in "theory" since 1810	
Personalization of issues; Proof by authority	Homeopathy = Hahnemann	
Lack of scientific method	No multiple controlled studies	

27

Identifying Pseudoscience - Software Engineering?

Use of vague, exaggerated or untestable claims	Dilutions of IC non-measurable; The higher the dilution, the better the effect; No rational underlying theory; Only proponents can "see" effect	"My technique makes programming easier" - How much easier? What does "easier" mean? How much would be "important"?
Over reliance on confirmation rather than refutation	Testimonials on effectiveness. No blind studies of effects	"I tried it and it works" - (ESEM audience is rather good at multiple studies)
Lack of openness to testing by other experts	Allowed to be sold in USA. FDA prevented by law from studying its effectiveness	"Paper is unpublishable since someone already ran that study"
Absence of progress	No change in "theory" since 1810	How often is a technique studied by a group other than the developer?
Personalization of issues; Proof by authority	Homeopathy = Hahnemann	Techniques often associated with developer. "I wrote it and I know what I'm doing."
Lack of scientific method	No multiple controlled studies	How often is falsification of results attempted?

28

Falsifiability

- The scientific method works by hypothesis generation followed by experimentation.
- A major goal of experimentation is falsifiability.
 - Philosopher Karl Popper asserted that a hypothesis, proposition, or theory is scientific only if it is falsifiable.
 - A major goal of experimentation is to show that the theory is false (e.g., the negative is false).
 - Only when this occurs can you begin to assert that maybe theory is correct.
- Have you read "the results do not confirm this theory, so we will modify the approach ..."

ESEM – October 2009

29

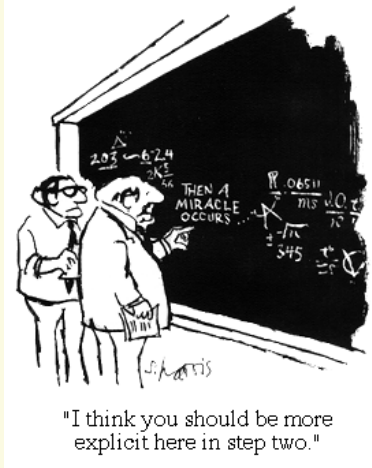
So how has software engineering been doing in the science domain?

- Where were we 25 years ago in applying the scientific method?
- Where are we today?
- What still has to be done?

ESEM – October 2009

30

The language of science - Mathematics



ESEM – October 2009

31

Need for measurement

- A quote you see quite often in experimental software engineering venues:
"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

- Corollary to above: **We need relevant measurements**

"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

ESEM – October 2009

32

Software engineering

- Lets start to get specific about SE.
- Does software engineering follow this model of science?
- In software engineering, the tools, methods, and techniques (e.g., the technologies) are our "theorems"
- Experimentation is how we partially validate (e.g., prove) whether those technologies are effective
 - Note: As an experimental approach toward proof, we can only give approximations as to how well each technology works
 - What can we say about experimentation in software engineering?

ESEM – October 2009

33

How do we measure the "science of software engineering"?

- Often there is a lack of validation before using a new technology
 - Anecdotal evidence that we don't validate our claims
 - Study by Tichy (1994) and Zelkowitz-Wallace (1998) confirm this
 - Only 15% of papers in other scientific fields
- Can we understand why this is so and how can we change this?

ESEM – October 2009

34

Role of experimentation studies

- Dolores Wallace and I reviewed over 600 published papers.
- Basic conclusion:
 - Approximately 50% of all software engineering papers had little validation of the claims in those papers.
 - Similar results were found by Walter Tichy in an earlier 1994 study.
- How does the software engineering community justify the many new technologies?
 - What methods are used to validate technologies?

ESEM – October 2009

35

Tichy study

- Data:
 - 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:
 - 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

ESEM – October 2009

36

1998 Validation methods

- Experimental models often taken from domains like psychology and medicine:
 - 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?
- This "classical method" of the controlled replicated experiment:
 - Not always feasible
 - Expensive (especially with large developments)
 - And there are other ways to evaluate technologies

ESEM – October 2009

37

Other experimental models

- Replicated experiments
 - Chemistry - Rows of test tubes
 - Psychology - Rows of freshmen students working on a task
- Observations of what happens
 - Medicine - Clinical trials, but "do no harm"
 - Astronomy - Observe events if and when they occur
- Data Mining of completed activities
 - Archaeology - Dig up the past
 - Forensic science - Recreate what happened

ESEM – October 2009

38

Developed a 15-step taxonomy of experimental methods

➤ Classes of methods

- **Controlled method** - Multiple instances of an observation in order to provide for statistical validity of the results.
- **Observational method** - Collect relevant data as it develops. In general, there is relatively little control over the development process.
- **Historical method** - Collect data from completed projects.

Basic 12-step program

1. *Project monitoring.* Collect accounting data from a project and then study it.
2. *Case study.* Collect detailed project data.
3. *Field study.* Monitor several projects (e.g., survey).
4. *Literature search.* Evaluate previously published studies.
5. *Legacy data.* Evaluate data from a previously-completed project.
6. *Lessons learned.* Perform a qualitative analysis on a completed project.
7. *Static analysis.* Use a control flow analysis tool on the completed project.
8. *Replicated experiment.* Develop multiple instances of a project.
9. *Synthetic.* Replicate a simpler version of the technology in a laboratory.
10. *Dynamic analysis.* Execute a program using actual data.
11. *Simulation.* Generate data randomly according to a theoretical distribution.
12. *Theoretical.* Formal description of an abstract model.

But the list is incomplete

- What 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.
 - We added this validation as a weak form of case study as an assertion.
- **Assertion** - A simple form of case study that does not meet scientific standards for experimental validation. More like a feasibility study than a validation.

ESEM – October 2009

41

Evaluation of taxonomy

- Do the 13 methods described previously make any sense?
- Do research groups really use them?
- This led to the 1998 Zelkowitz-Wallace study and the 2006 update to it
(IEEE Computer, May 1998)

ESEM – October 2009

42

1998 study

- National Institute of Standards and Technology
- Validate Tichy conclusions on a wider sample
 - Can we classify methods used to validate claims in a larger context than the Tichy survey?
- Looked at papers published in 1985, 1990, 1995
 - Sources - All papers published in those years in:
 - IEEE Software
 - Transactions on Software Engineering
 - ICSE proceedings
 - 612 papers reviewed
- Every paper classified by 2 people
- But 13 categories were not enough

ESEM – October 2009

43

Non-validation methods

Validation methods

12 validation methods given previously

Non-validation methods

- **Assertion.** Informal feasibility demonstration.
- **Not applicable.** Paper was not appropriate for an experimental validation (e.g., tutorial, survey, news item)
- **No experimentation.** Default if none of the previous 14 methods applied. (e.g., paper should have had a validation, but didn't)

ESEM – October 2009

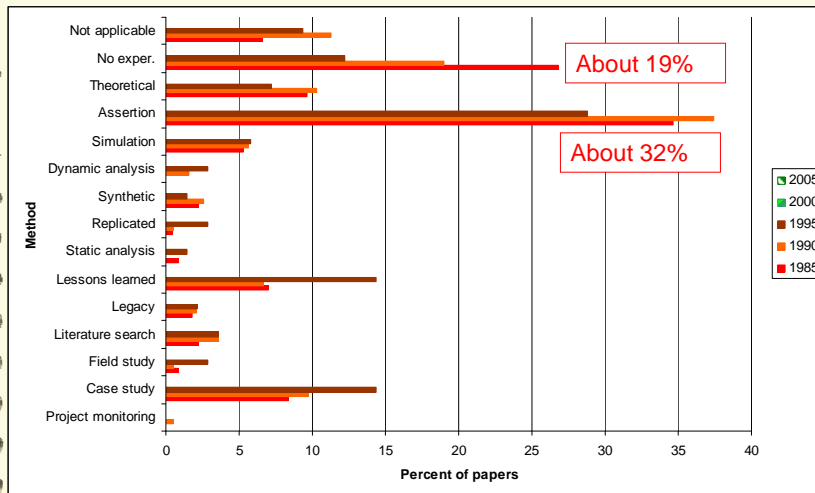
44

Basic Data (1998 study + 2006 update)

	Proj Mon	Case study	Field Study	Lit Sch	Leg	Less Lrn	Stat Anal	Rep	Syn	Dyn Anal	Sim	Ass	The	No Exp	NA	TTL
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	7	3	1	1	3	0	0	3	0	0	4	1	11	19	53
2000 Total	0	26	6	2	2	7	0	2	9	8	8	26	14	46	31	187
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	9	4	1	5	0	2	1	2	13	5	13	1	8	2	66
2005 Total	0	32	9	2	11	0	4	2	7	34	11	36	3	19	4	174

ESLMI - October 2009

1998 data



Conclusions from 1998 study

- 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
- BUT
 - Almost 20% of the papers had no experimental validation
 - Assertions (a weak form of validation) were about one-third of the papers
 - Resulting in over 50% of the papers having no real validation! (Different methodology, but same basic result as Tichy survey.)

Unexpected conclusions from 1998 study

- Every paper could fit into one of our categories, but:
 - Some papers can apply to 2 categories. We chose what we believed to be the major evaluation category.
 - We ignored what author said they were doing and tried to figure it out from context.
- Sometimes category extremely hard to uncover.
 - Authors often never stated why they were writing this paper
 - e.g., "In this paper we describe our database Zork" never once saying why we need another database product or what it will do for us that existing database products don't do.
 - Authors fail to state how they propose to validate their hypotheses
 - Words like experiment, prototype, validate, pilot study, case study, ... used to mean many different things. We ignored those words and tried to objectively classify paper

But there was one interesting result

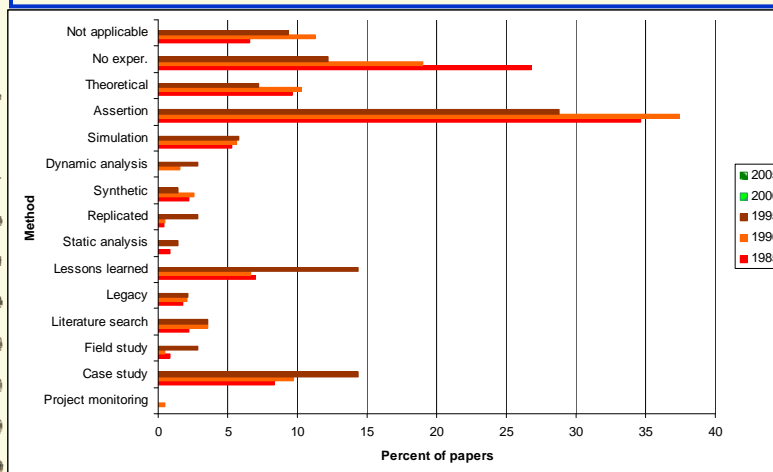
- But percentages of "no experimentation" dropped from 27% in 1985 to 19% in 1990 to only 12% in 1995.
 - Perhaps indicative of a favorable trend
- By 2006, 2 more data points available (2000 and 2005).
 - Perhaps a revised survey would show something interesting
 - So survey extended in 2006: 361 additional papers classified

Zelkowitz M. V., An update to experimental models for validating computer technology, *Journal of Systems and Software* 82, 2009, 373-376

ESEM – October 2009

49

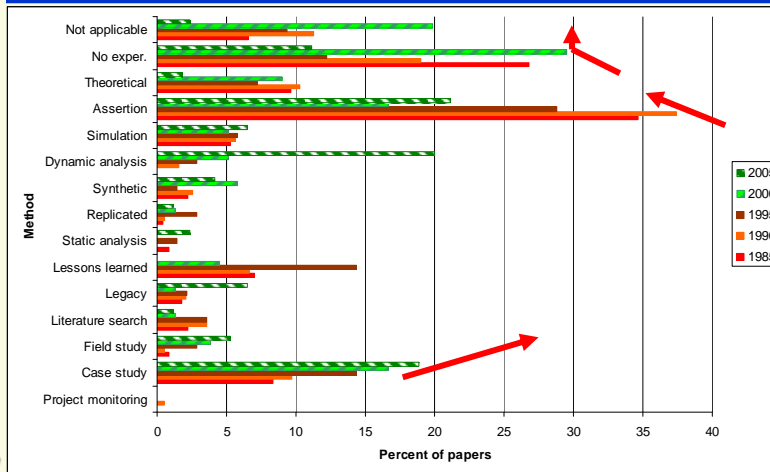
1998 data



ESEM – October 2009

50

1998 + 2006 data



ESEM – October 2009

51

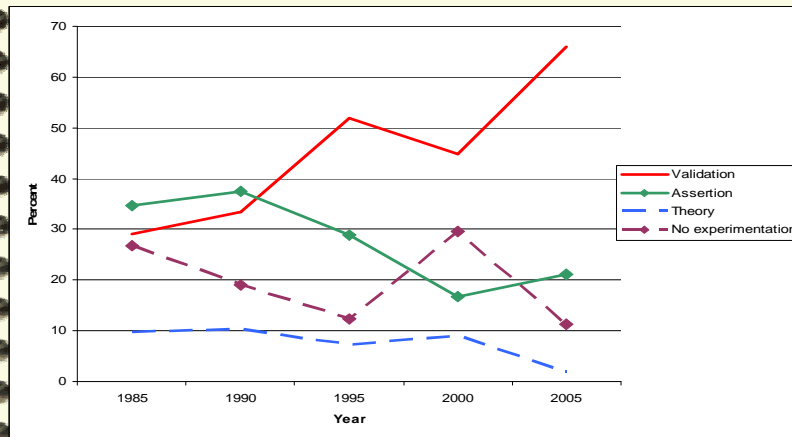
Anecdotal information

- "Case studies" steadily increasing
- "Assertions" dropping
- "No experimentation" dropping - generally, but not much
- More use of data repositories
 - Open source development histories (Mozilla, Apache) source of many papers
 - Increase in "Legacy" and "Dynamic analysis" methods
- Little change in controlled experiments (small increase to 7% of total)

ESEM – October 2009

52

Trends



ESEM – October 2009

53

Threats to validity

- Consistency of taxonomy process
- Each data point very dependent on specific editors and program committees (e.g., 2000 ICSE)
- Change of scope in IEEE Software
- **Quality** of validation not indicated
 - Only tried to classify method used to validate paper, not whether the validation was correct

ESEM – October 2009

54

Study results

- We have proposed a 15-way approach toward developing a quantitative model of software experimentation.
- In general, the trend observed in 1998 that an increasing number of papers have an empirical validation component seems to be continuing through 2005.
- Informal feasibility studies (assertions) seem to be greatly declining
- Still need to look at the quality of those evaluations

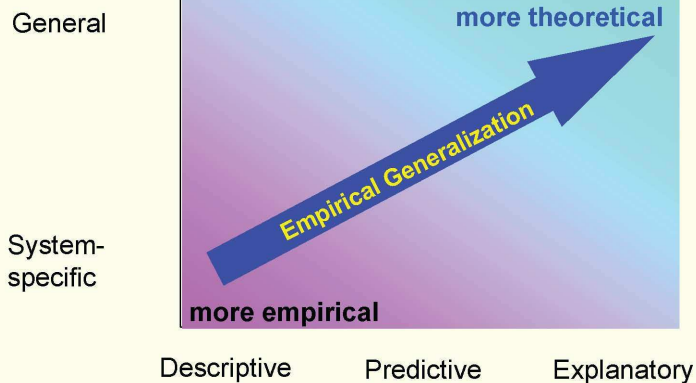
Where do we go from here????

- Experimentation is only the beginning
- We need theories to explain software engineering phenomena

Aspects of a scientific theory

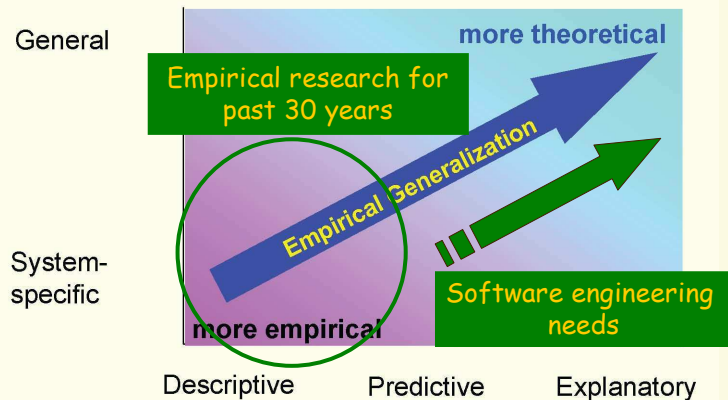
- **Parsimony:** Explains a variety of phenomena with a short, coherent explanation (*Occam's razor*).
- **Generality:** Permits a reduction in the number of independent phenomena, with the single theory explaining a wide range of phenomena.
- **Prediction:** Anticipates future events, often to a high degree of accuracy.
- **Explanatory:** Provides a compelling underlying mechanism.

Development of theories



From P. Cohen, *Empirical Methods for Artificial Intelligence*, 1995

Development of theories



From P. Cohen, *Empirical Methods for Artificial Intelligence*, 1995

ESEM – October 2009

59

There are activities looking at the problem

- R. Snodgrass- U. of Arizona - Ergalics
- Discussion of role of conferences and journals in *Comm. of the ACM* (May 2009)
- Data sharing, ISERN meeting, 2005
- Empirical Software Engineering Journal
- But few community-wide discussion of the problems

ESEM – October 2009

60

Summary

- We need to become better skeptics, both outside and within the software engineering community.
- We need to be more aware of when we deviate into voodoo science and must avoid the pitfalls of pseudoscience.
- We have few predictive models and almost no explanatory models of software processes - We cannot tell researchers "invent new better theories", but we can be more aware of that goal.
- Although computer science has vastly improved its adherence to the scientific method over the past 25 years, we still can do better.
- **Take-away message:** Until we overcome those limitations, our impact on the general scientific establishment will be limited.

ESEM – October 2009


61

That's all folks!



ESEM – October 2009

62

The image shows a spiral-bound notebook with a light brown, textured cover. The spiral binding is on the left side. The text is centered on the cover.

The Science of Software Engineering

Marvin Zelkowitz
Department of Computer Science
University of Maryland
College Park, Maryland
marv@zelkowitz.com

ESEM - 2009

ESEM - October 2009

63