In God we trust. Everyone else, bring data.

Michael Bloomberg, NYC Mayor

DTU

Prof. Dr. Harald Störrle
Danmarks Tekniske Universitet

Alternative Ways to Convince People

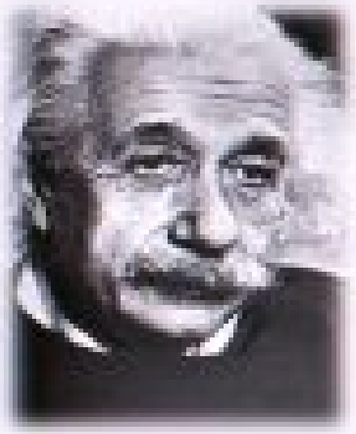
Alternative Ways to Convince People

By force

physical force, group pressure

Alternative Ways to Convince People

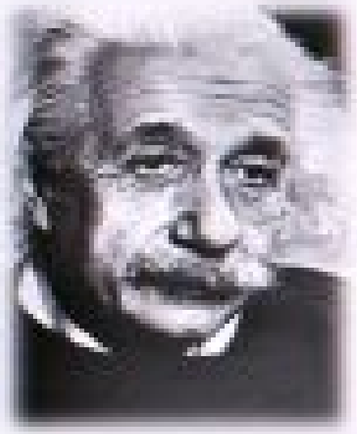
By force
physical force, group pressure



By authority
divine revelation, fame

Alternative Ways to Convince People

By force
physical force, group pressure

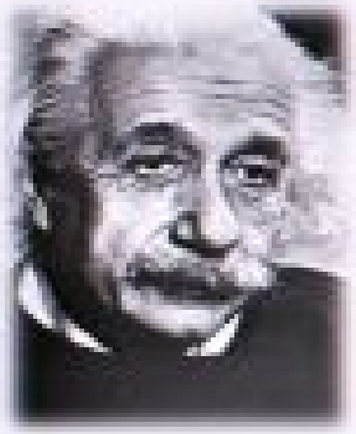


By authority
divine revelation, fame

By insight
plausibility, observation

Alternative Ways to Convince People

By force
physical force, group pressure



By authority
divine revelation, fame

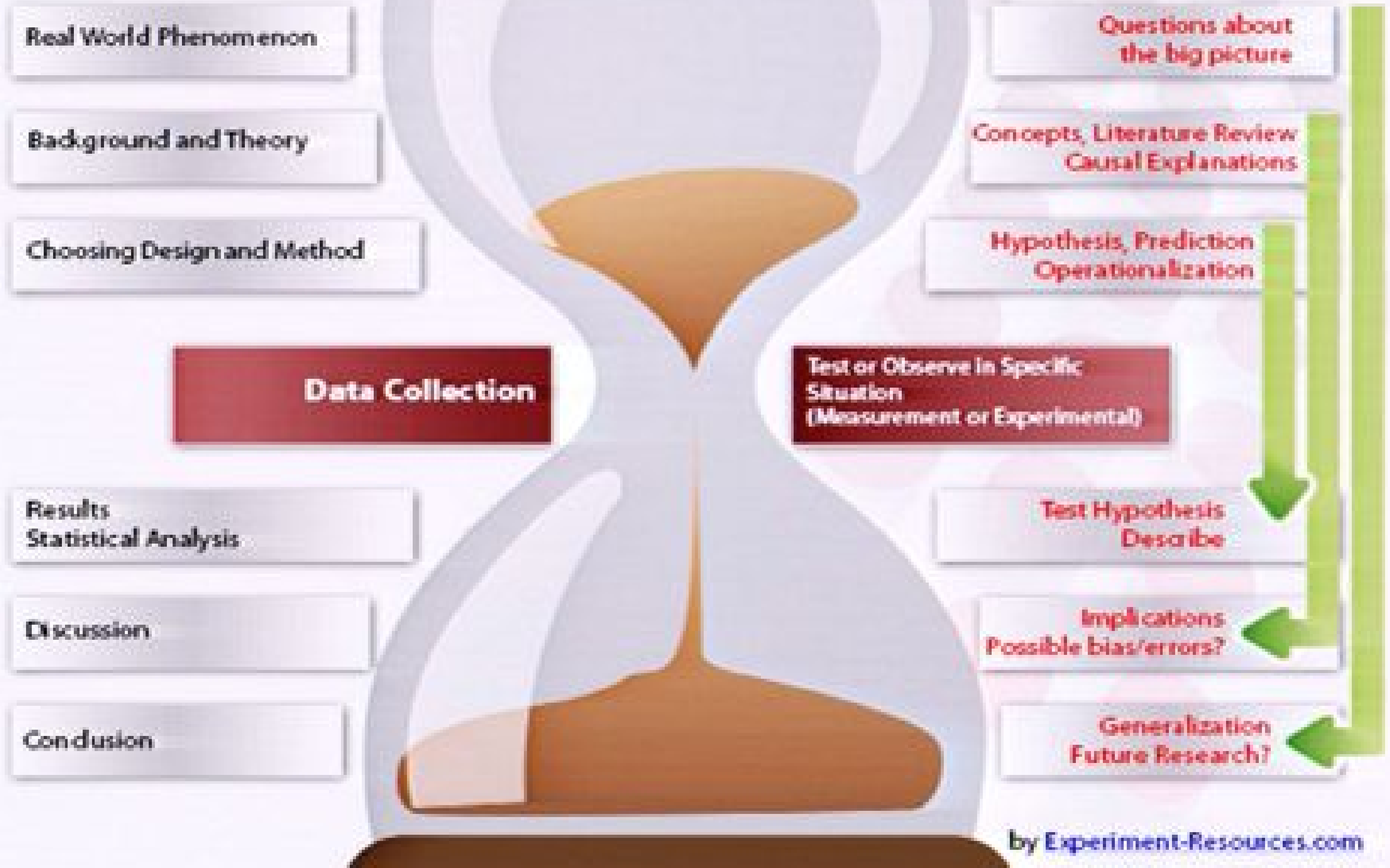
By insight
plausibility, observation

By research

The Process of Scientific Research

by Experiment-Resources.com

The Process of Scientific Research

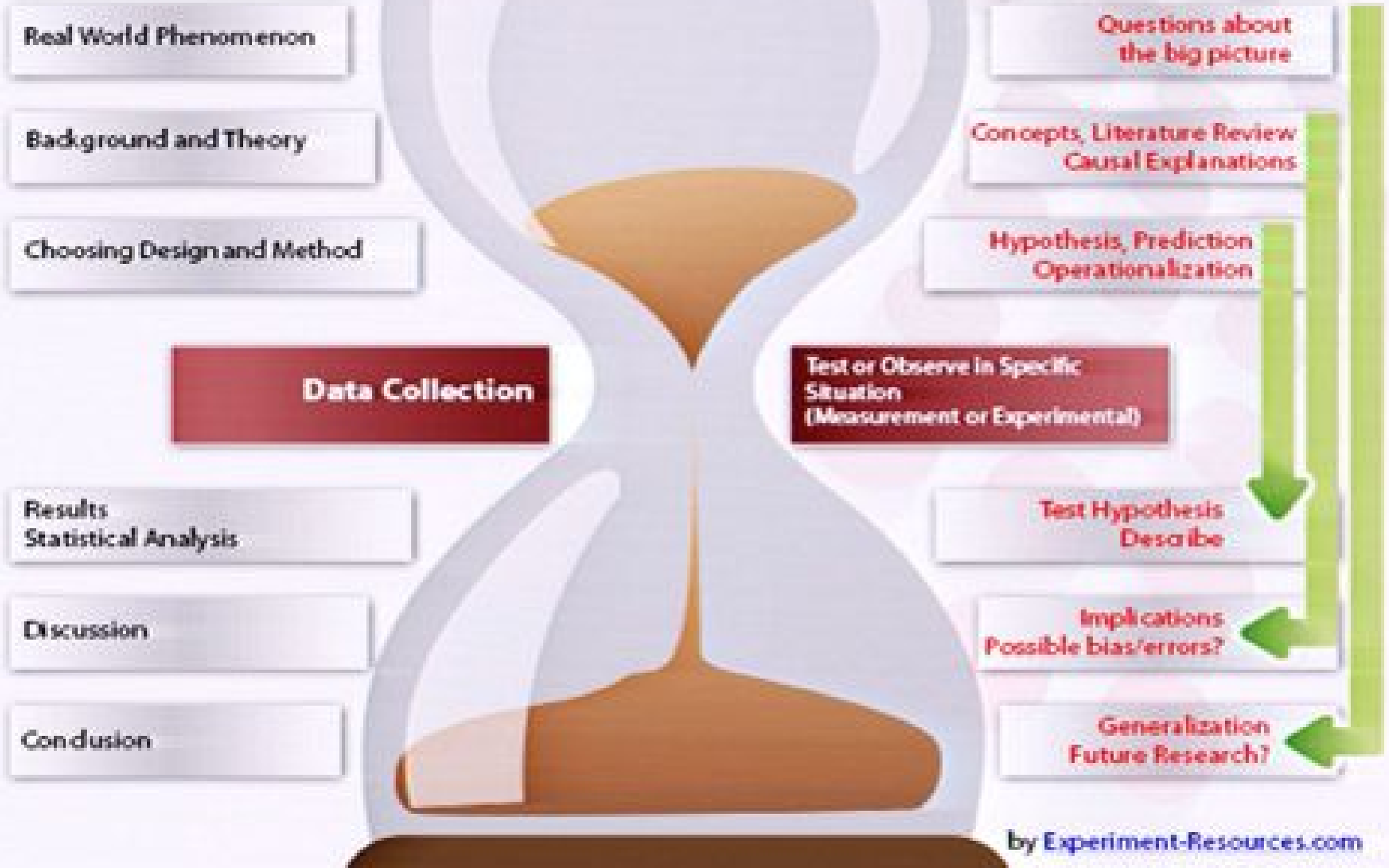


by Experiment-Resources.com

The Process of Scientific Research

by Experiment-Resources.com

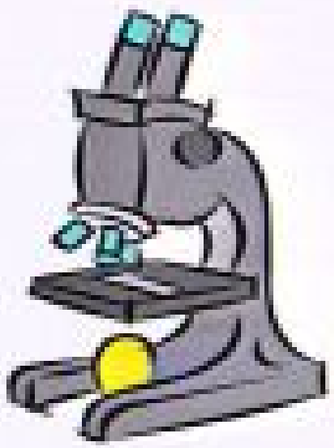
The Process of Scientific Research



by Experiment-Resources.com

Empirical Methods in Sciences

- **In the natural and social sciences, there are three major classes of research instruments:**
  - Controlled experiments
  - Case Studies and Field Studies
  - Surveys, Structured Reviews, and Polls
- **They all have their respective strengths and weaknesses.**



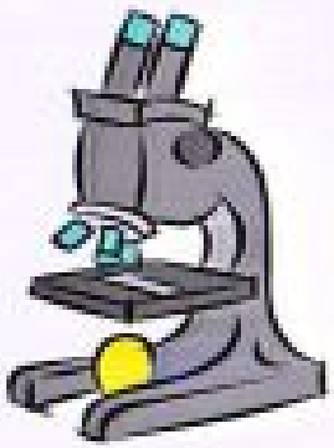
Controlled Experiment

Case Study

Survey / SLR

Empirical Methods in Sciences

- **In the natural and social sciences, there are three major classes of research instruments:**
  - Controlled experiments
  - Case Studies and Field Studies
  - Surveys, Structured Reviews, and Polls
- **They all have their respective strengths and weaknesses.**



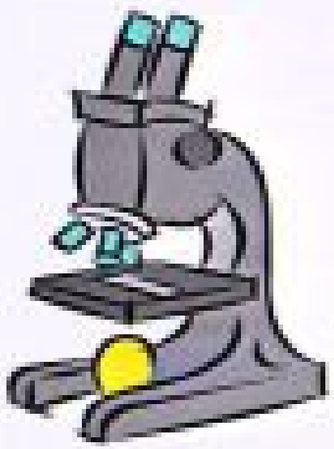
Controlled Experiment

Case Study

Survey / SLR

Empirical Methods in Sciences

- **In the natural and social sciences, there are three major classes of research instruments:**
  - Controlled experiments
  - Case Studies and Field Studies
  - Surveys, Structured Reviews, and Polls
- **They all have their respective strengths and weaknesses.**



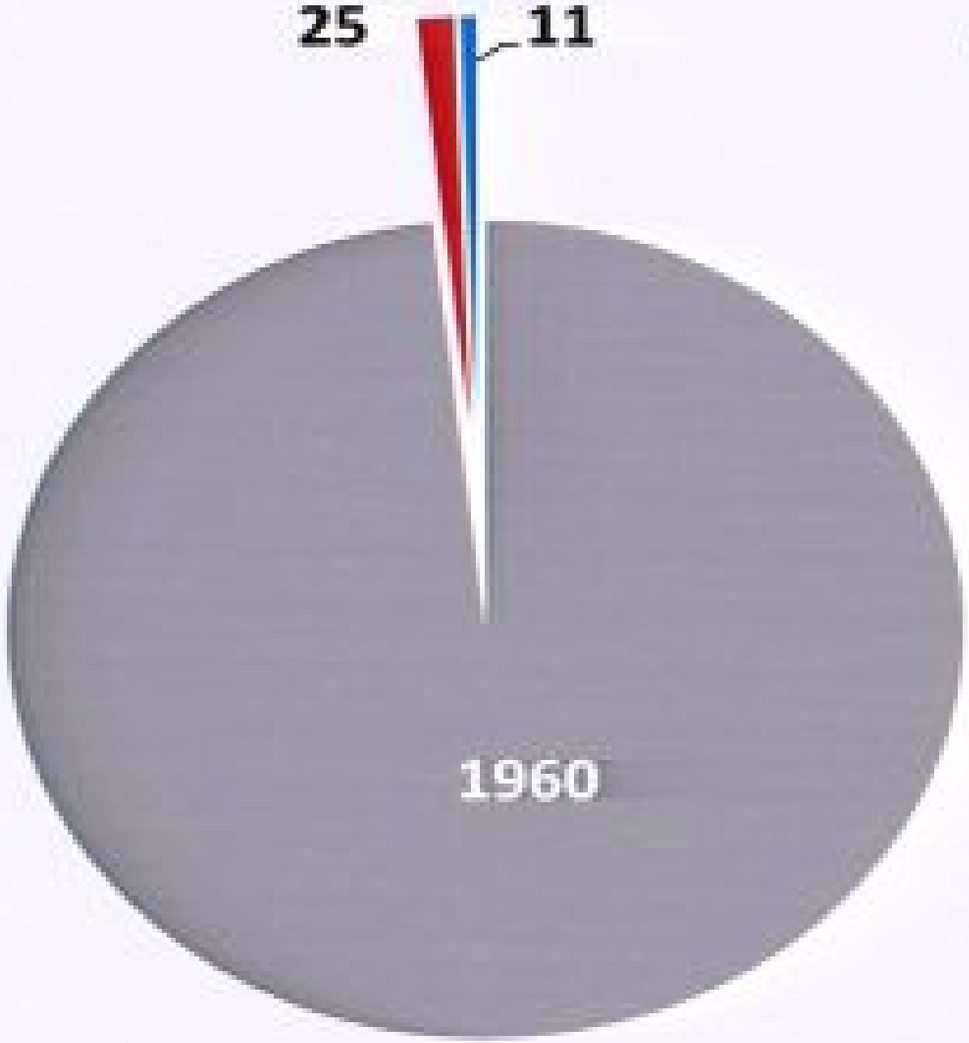
Controlled Experiment

Case Study

Survey / SLR

What do we know about Agile SE?

Meta-review (SLR) of all published studies relating to agile practices up to 2005.



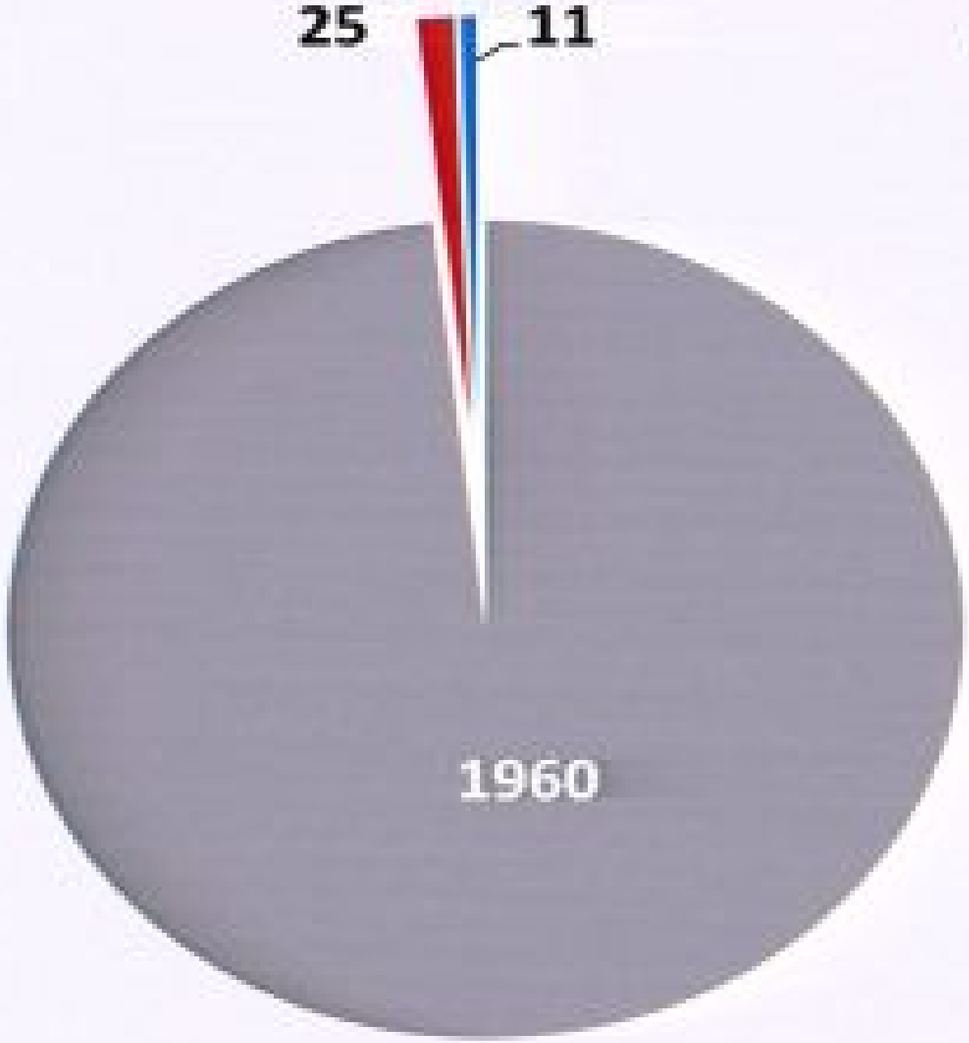
All studies on agile approaches

Empirical Studies (XP)

Empirical Studies (Other)

What do we know about Agile SE?

Meta-review (SLR) of all published studies relating to agile practices up to 2005.

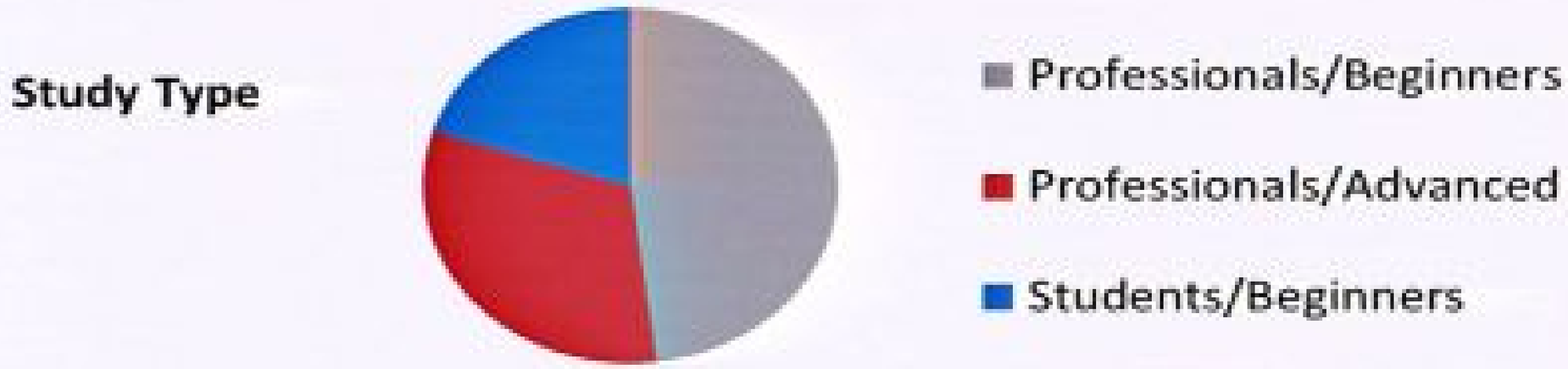


All studies on agile approaches

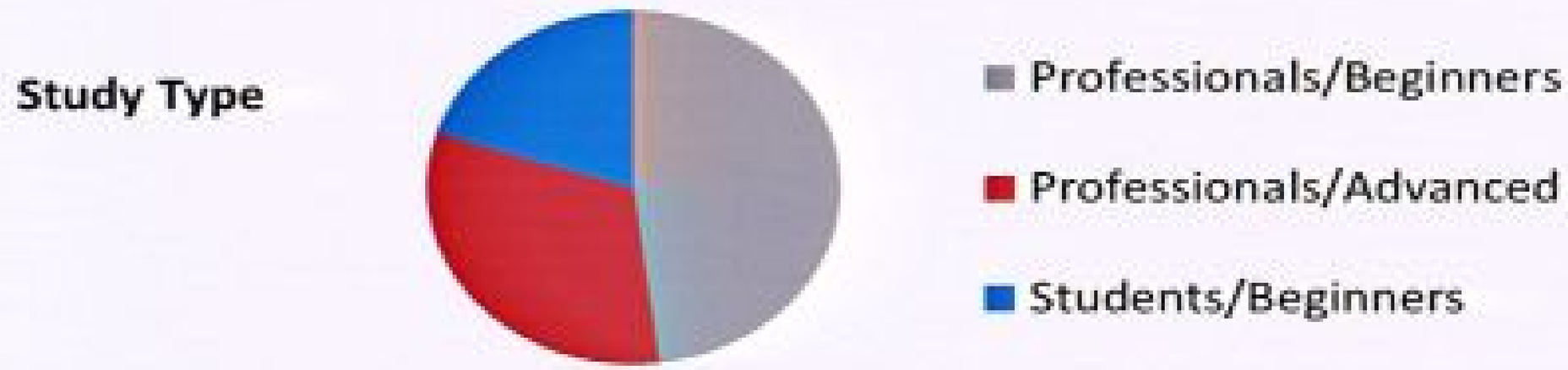
Empirical Studies (XP)

Empirical Studies (Other)

How reliable is our knowledge?

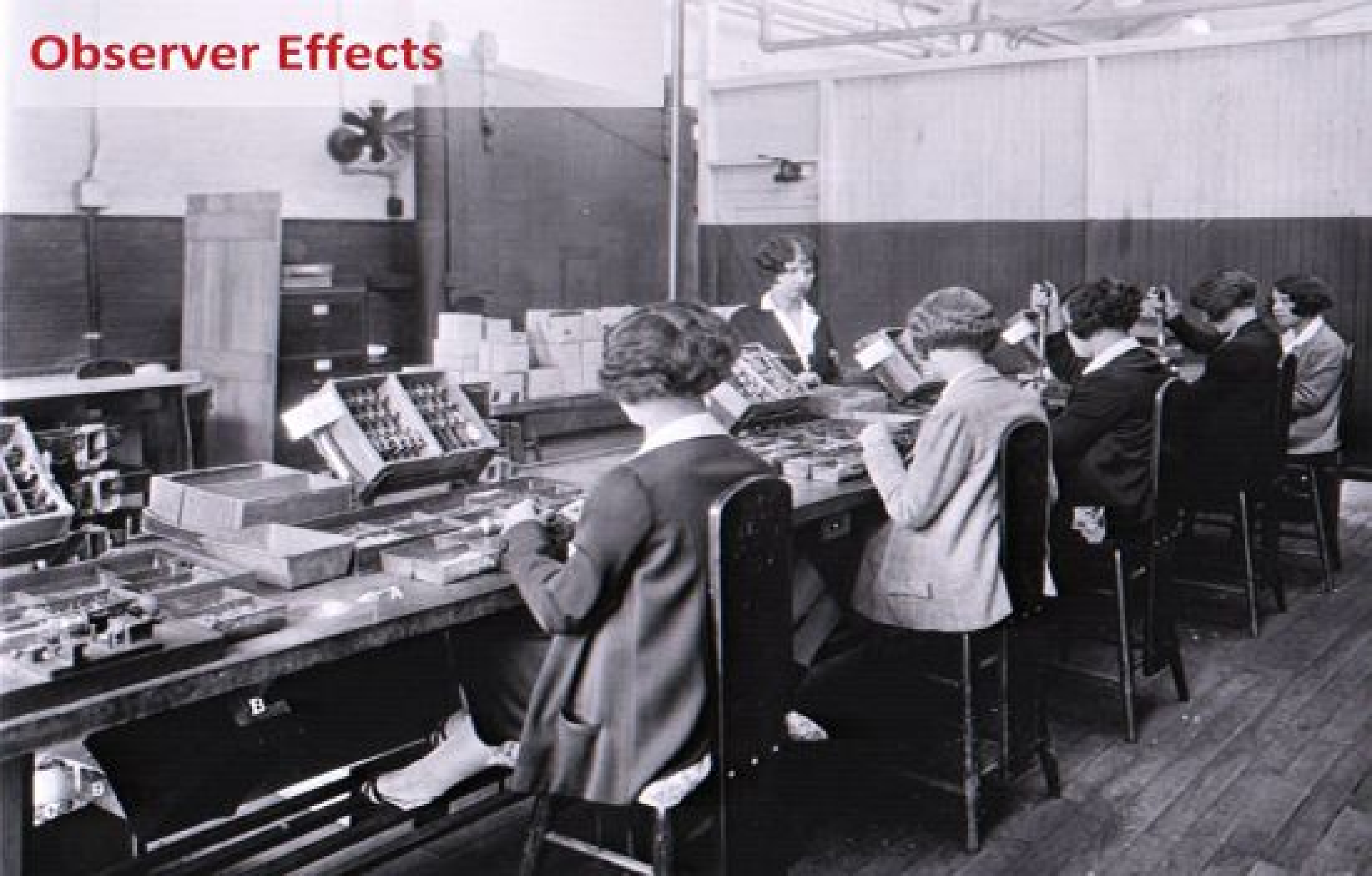


How reliable is our knowledge?



The Critical Appraisals Skill Program (CASP) is a checklist to asses rigor, credibility, and relevance of empirical research, in particular those using qualitative methods.

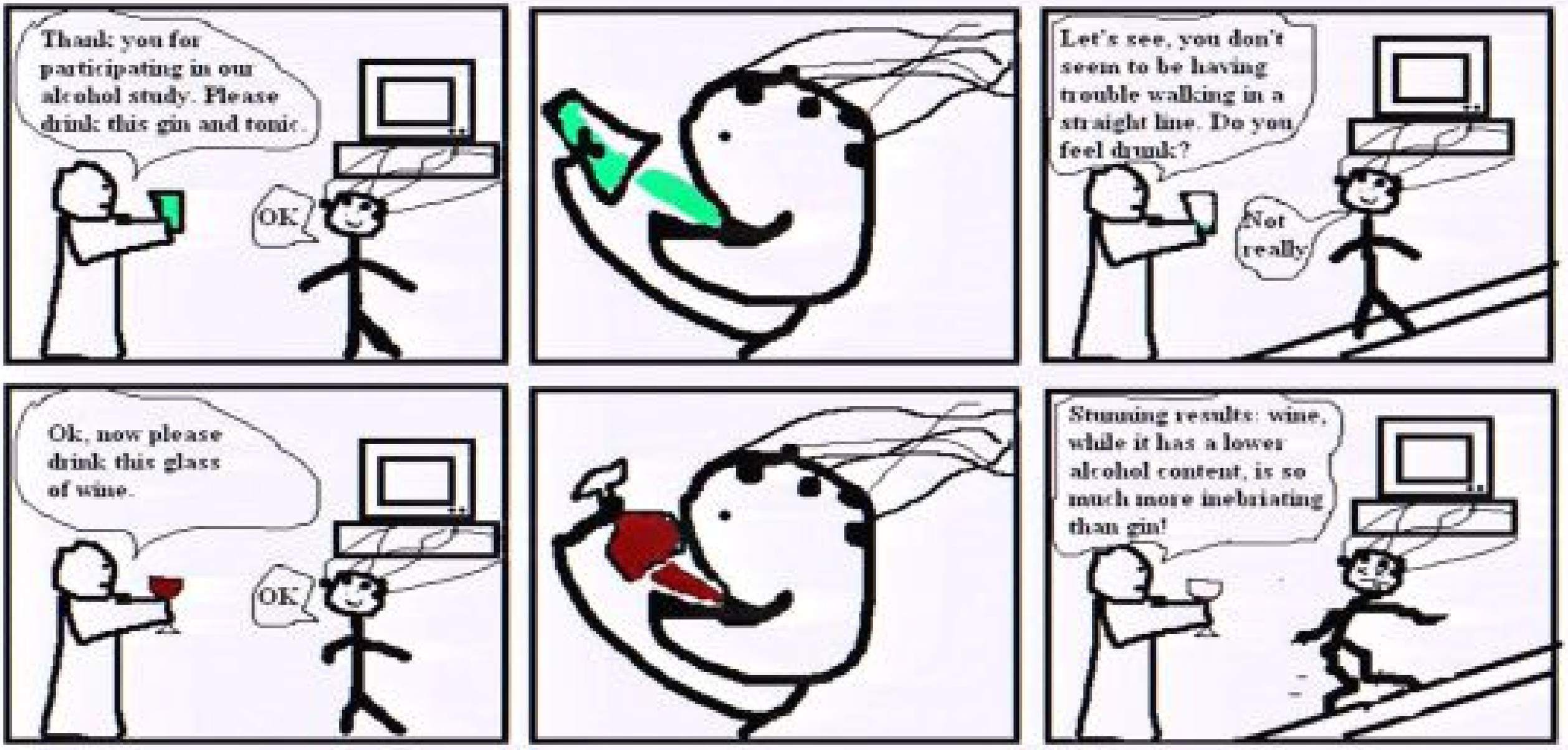
Observer Effects



Anecdotes are not scientific evidence

The highly respected Professor Nibbowitz proved,
that octopus are more intelligent than cat,
when exposed to the same challenges and conditions.

Learning Effects



Evidence for MBSD

Evidence for MBSD

- **What is the amount and quality of the evidence?**
  - There are plenty success stories from industry, many surveys on modeling practices in industry, and there are also large numbers of experiments published.
  - There is no SLR on (economic) effects of MBSD, but there are two recent series of qualitative studies to investigate this question, both of which are considered methodologically sound.
  - Also, there are several meta-studies on UML, and model quality.

- **The state of scientific knowledge about Model Based Sw. Development is a little bit better than for Agile Sw. Development.**
  - This might be due to the fact that "Modeling" became fashionable earlier than "Agile".
  - Also, being closer to academia than to industry, scientific validity was more of a concern to the key players.
  - Finally, the knowledge we have about MBSD comes mostly from qualitative methods which are by definition more suitable for answering "large" questions than the experimental paradigm that dominates in the ASD world.

Some results about MBSD

Some results about MBSD

- **The benefit of using MBSD in industry is not necessarily a large speedup/efficiency gain, but higher system quality, better maintainability, and smaller systems.**

Some results about MBSD

- **The benefit of using MBSD in industry is not necessarily a large speedup/efficiency gain, but higher system quality, better maintainability, and smaller systems.**

- **MBSD is more useful for larger projects/companies.**

- **Generating code/documents from models is cost-effective in projects with high criticality or heavy regulatory constraints.**

Some results about MBSD

- **The benefit of using MBSD in industry is not necessarily a large speedup/efficiency gain, but higher system quality, better maintainability, and smaller systems.**

- **MBSD is more useful for larger projects/companies.**

- **Generating code/documents from models is cost-effective in projects with high criticality or heavy regulatory constraints.**

- **MBSD benefits are, to a large degree benefits derived from modeling as such.**

- **Visual Languages are not necessarily better than textual ones.**

Some results about MBSD

- The benefit of using MBSD in industry is not necessarily a large speedup/efficiency gain, but higher system quality, better maintainability, and smaller systems.

- MBSD is more useful for larger projects/companies.

- Generating code/documents from models is cost-effective in projects with high criticality or heavy regulatory constraints.

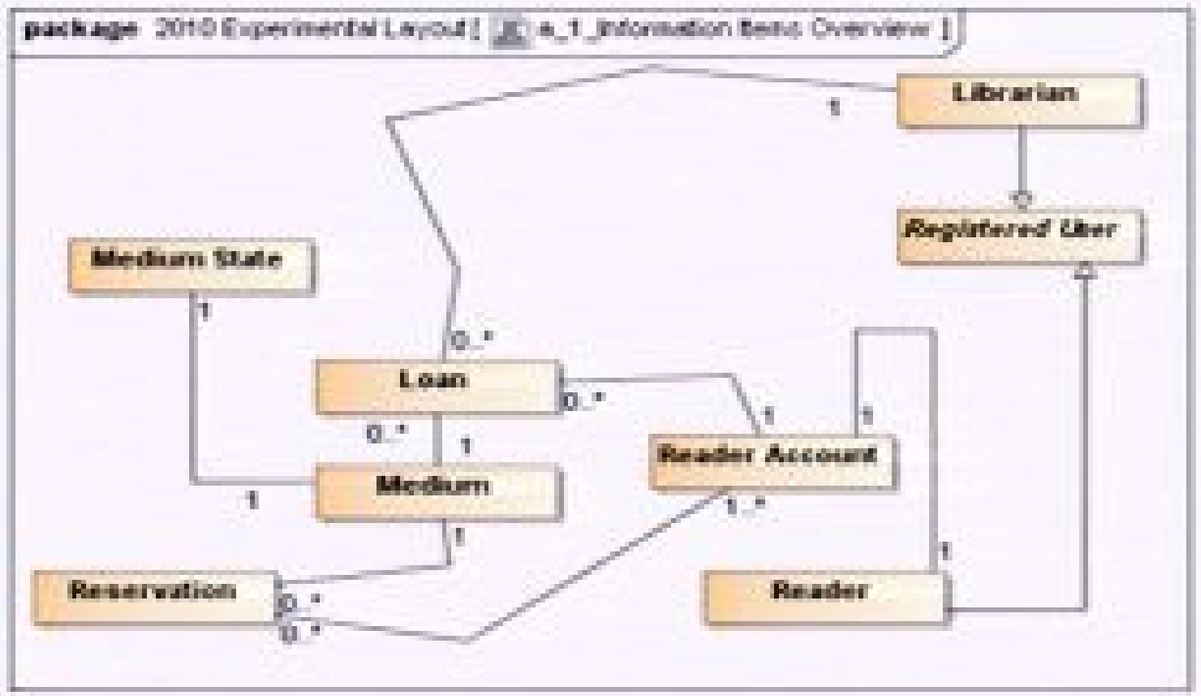
- MBSD benefits are, to a large degree benefits derived from modeling as such.

- Visual Languages are not necessarily better than textual ones.

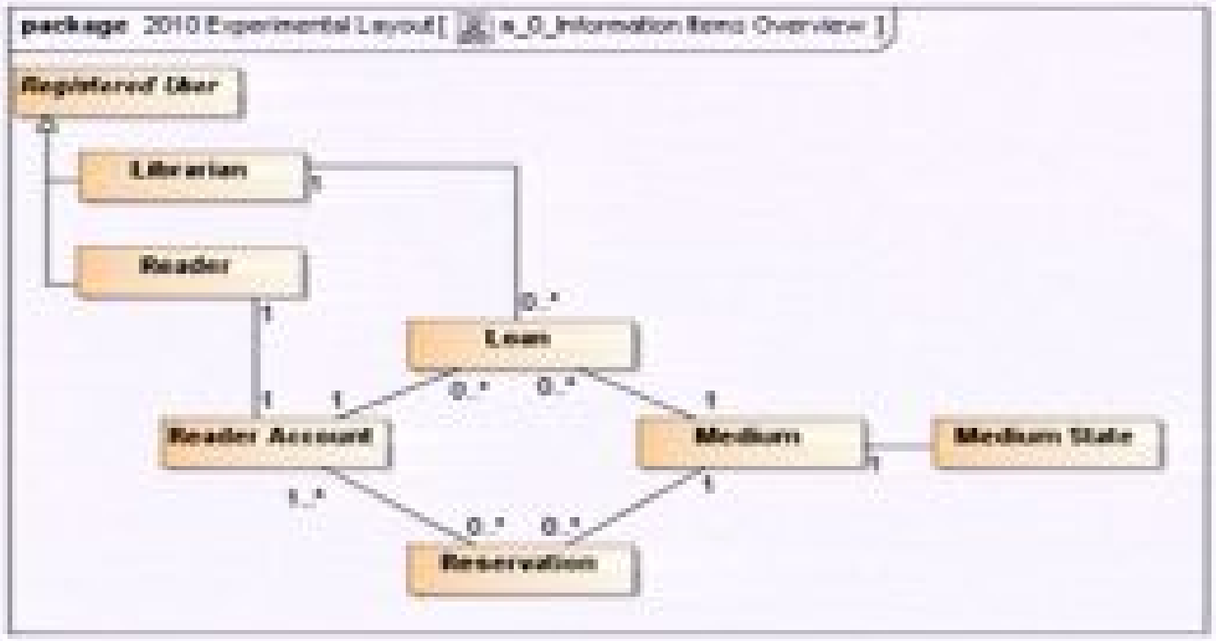
- Diagram layout is critical for model understanding.

Good vs. Bad (UML Diagram) Layout

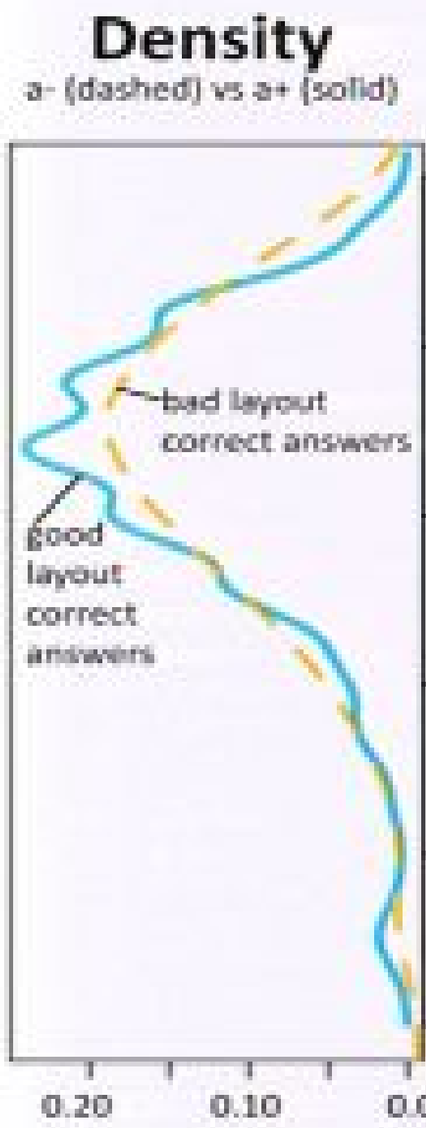
- **Elements of bad layout**
  - Edge crossings and bends
  - Overlaping/obscuring elements
  - Varying colors/sizes
  - Varying text orientation

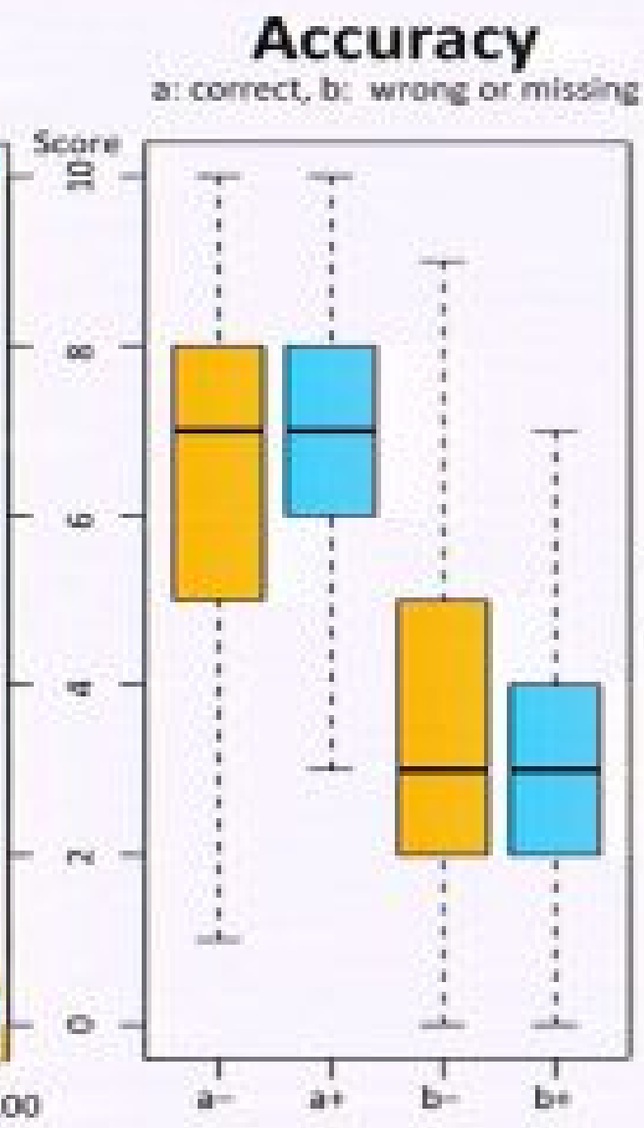


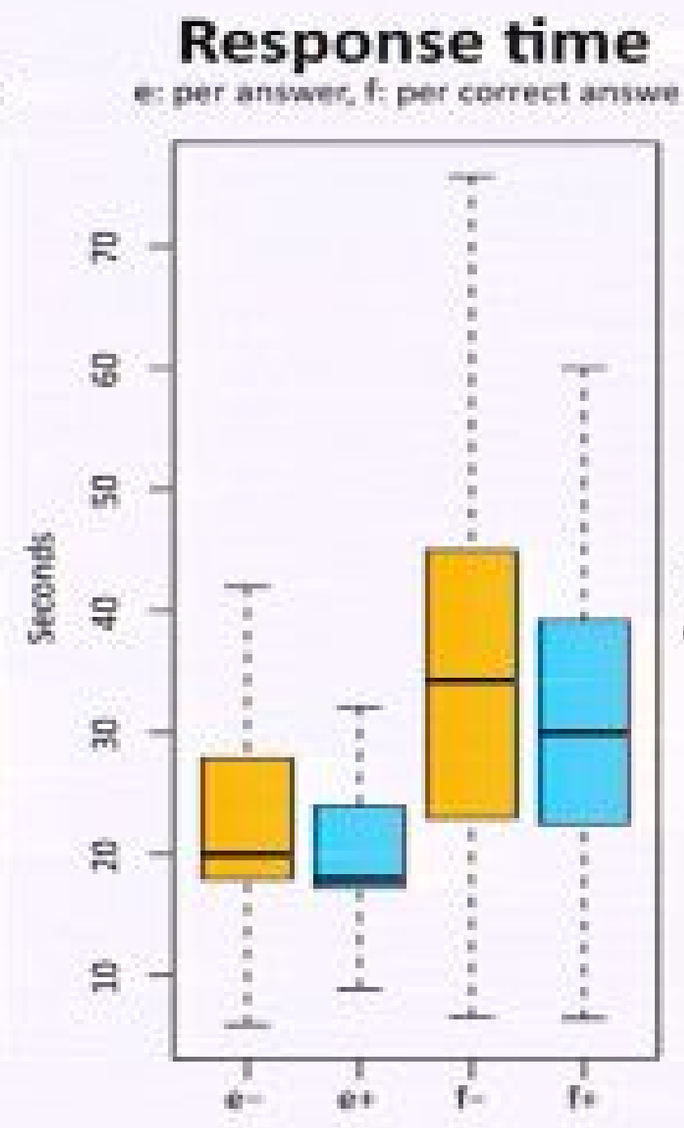
- **Elements of good layout**
  - Join similar edges
  - Cluster similar elements
  - Orthogonal arrangement
  - Place elements in flow

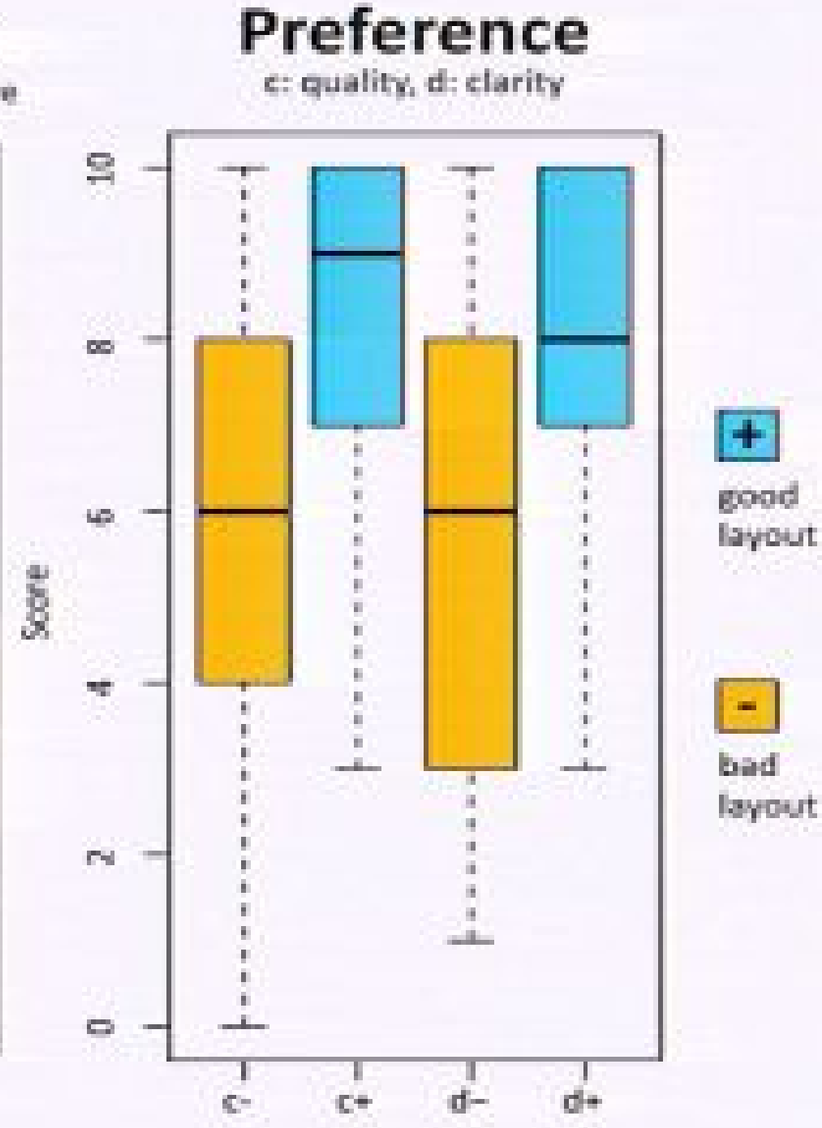


Observations









Effect Size

- **In contrast to previous studies, we observe a comparatively large effect, though not necessarily in user performance.**

Accuracy (a, b)

answers	bad layout μ_b	bad layout σ	good layout μ_g	good layout σ	benefit $\mu_g - \mu_b$
right	6.35	2.07	6.76	1.94	+6.5%
wrong/missing	3.65	2.07	3.24	1.94	-12.7%

Preference (c, d)

rating	bad layout μ_b	bad layout σ	good layout μ_g	good layout σ	benefit $\mu_g - \mu_b$
diagram quality	5.54	2.74	8.06	2.12	+31.3%
diagram clarity	5.61	2.74	7.81	2.27	+28.2%

Response time (e, f)

s/answer	bad layout μ_b	bad layout σ	good layout μ_g	good layout σ	benefit $\mu_g - \mu_b$
all answers	22.72	10.85	21.06	8.25	-7.3%
right answers	38.37	24.39	31.68	15.77	-17.4%

- **Cognitive load seems to benefit much more from good layout than objective performance indicators.**
  - This might be due to subjective coping strategies.
  - Dual stimulus experiments might shed light on this hypothesis.

Expertise ~ Impact

Initially, we found no noticeable differences between novice modelers and advanced modelers, to our surprise.

Maybe, our "experts" were no real experts?

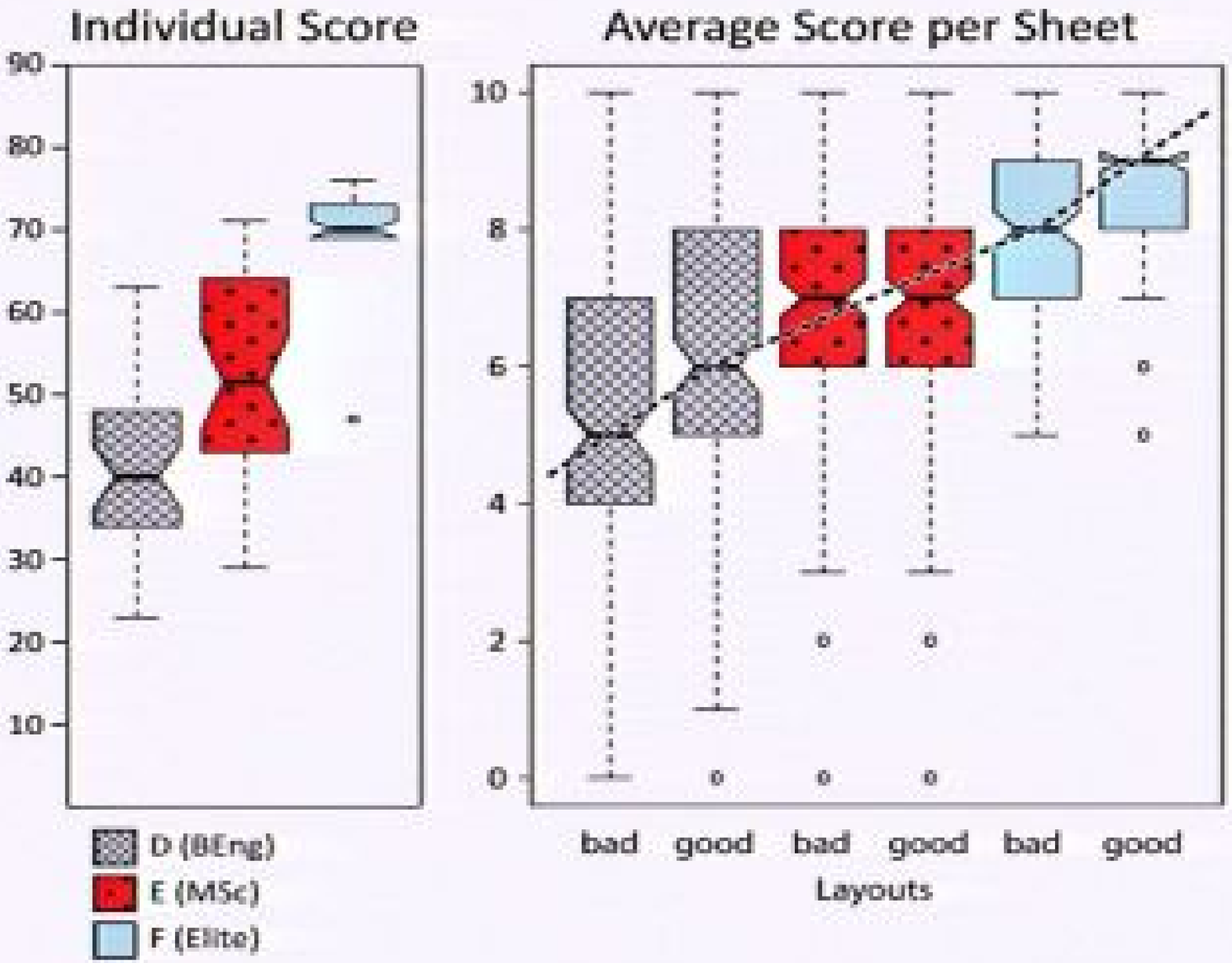


Diagram Size ~ Impact

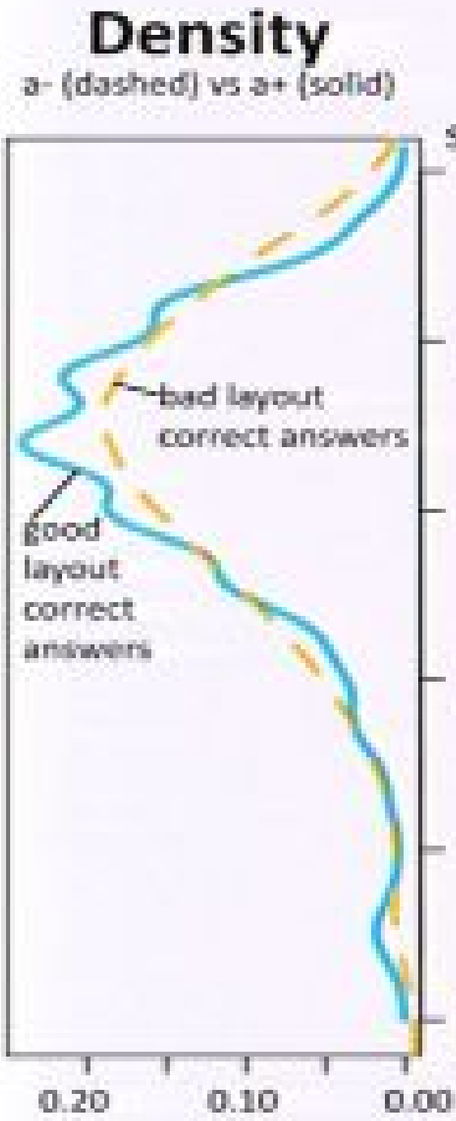
Objective Performance	Score Mean				Score Variance			
	r	ES	p	SIG	r	ES	p	SIG
All Diagrams	−0.423	L	0.010	**	0.424	L	0.010	**
Bad Layout	−0.491	L	0.039	*	0.534	L	0.023	*
Good Layout	−0.396	M	0.104	*	0.303	M	0.222	

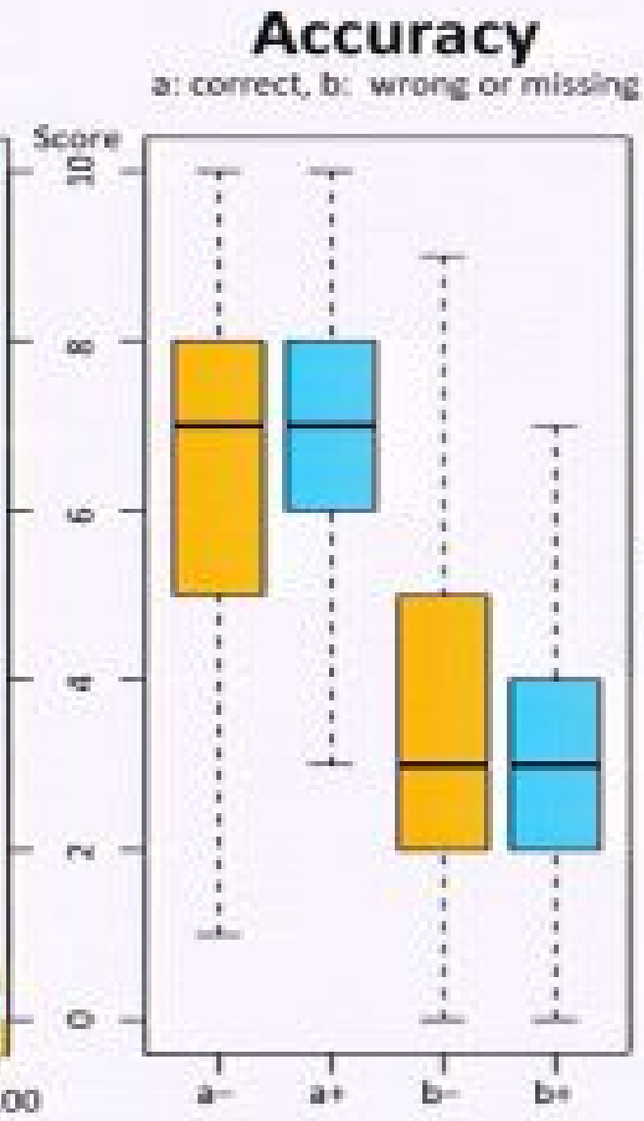
Diagram Assessment	Layout Quality				Layout Clarity			
	r	ES	p	SIG	r	ES	p	SIG
All Diagrams	0.538	L	< 0.001	***	−0.508	L	0.002	**
Bad Layout	0.521	L	0.027	*	−0.563	L	0.015	*
Good Layout	0.573	L	0.013	*	−0.766	L	0.0002	***

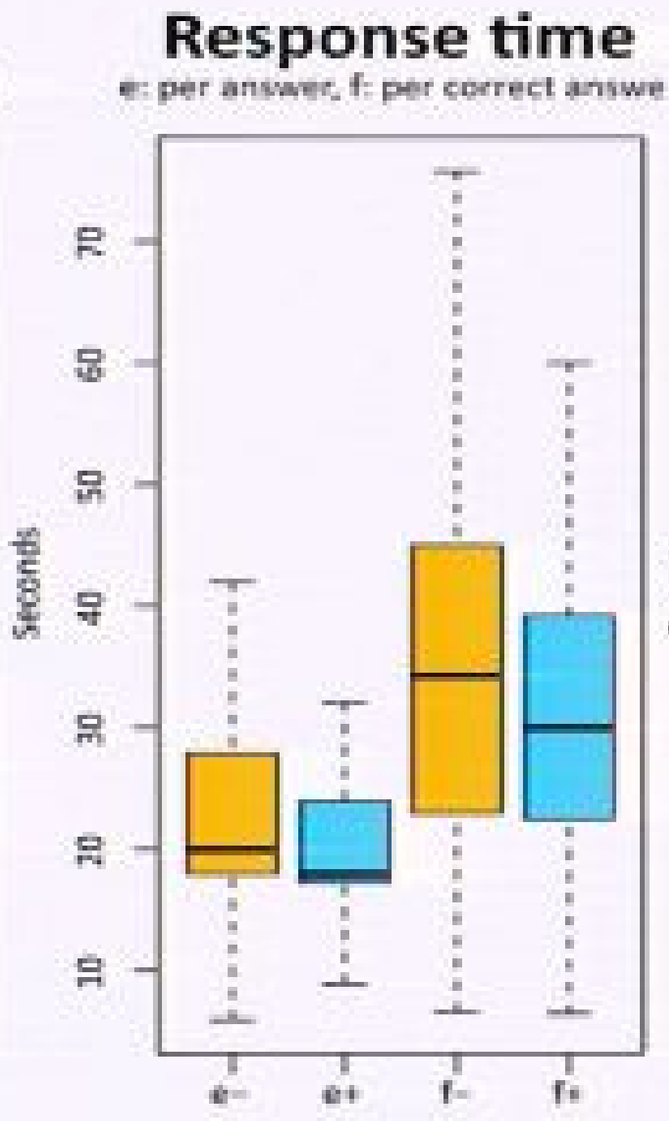
Cognitive Load	Diagram Understanding				Diagram Complexity			
	r	ES	p	SIG	r	ES	p	SIG
All Diagrams	−0.338	M	0.044	**	−0.081	S	0.640	
Bad Layout	−0.452	L	0.060	*	−0.313	M	0.207	
Good Layout	−0.197	S	0.434		0.152	S	0.548	

Table 2. Pearson's product-moment correlation between diagram size and modeler performance, measured as mean and variance of objective performance (correct answers, i.e., score), different subjective assessments, and cognitive load measures. In each cell, the first number is Pearsons's r indicating the size of the the correlation, the letter S/M/L classifies the effect size, the next number is the p-value, and the stars indicate its significance level.

Observations







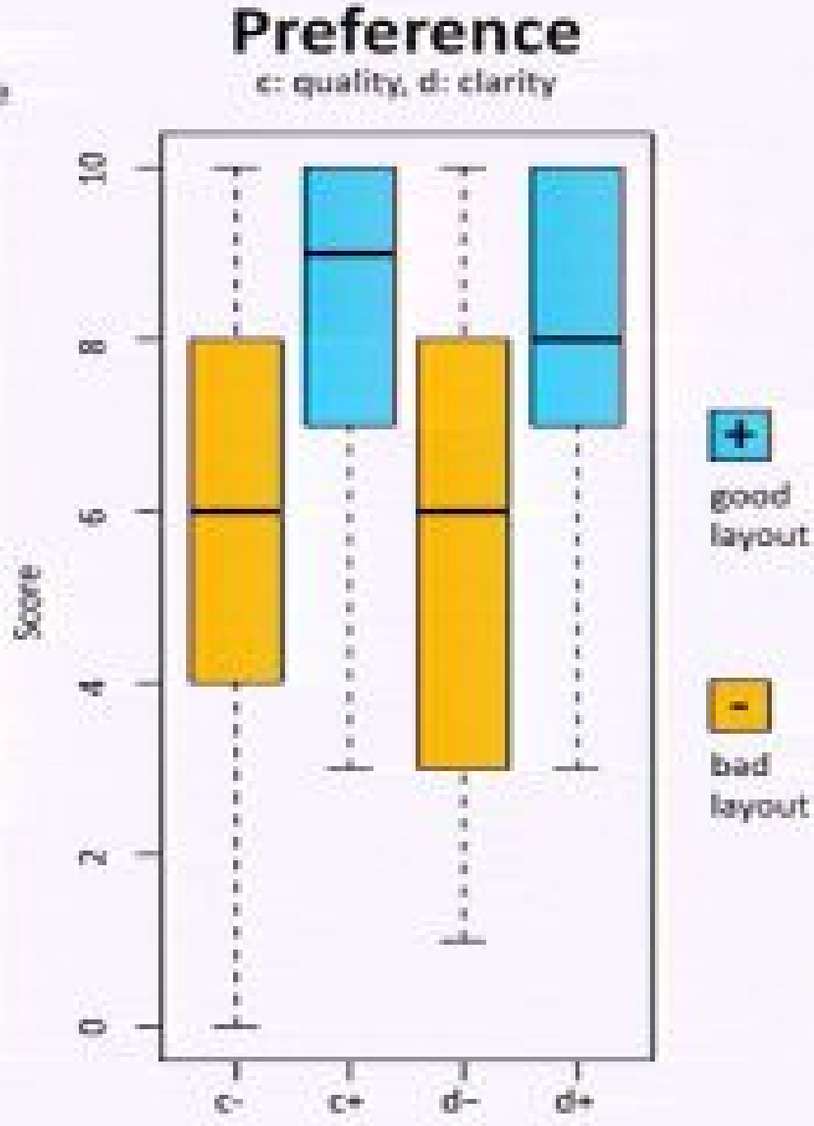


Diagram Size ~ Impact

Objective Performance	Score Mean				Score Variance			
	r	ES	p	SIG	r	ES	p	SIG
All Diagrams	−0.423	L	0.010	**	0.424	L	0.010	**
Bad Layout	−0.491	L	0.039	*	0.534	L	0.023	*
Good Layout	−0.396	M	0.104	*	0.303	M	0.222	

Diagram Assessment	Layout Quality				Layout Clarity			
	r	ES	p	SIG	r	ES	p	SIG
All Diagrams	0.538	L	< 0.001	***	−0.508	L	0.002	**
Bad Layout	0.521	L	0.027	*	−0.563	L	0.015	*
Good Layout	0.573	L	0.013	*	−0.766	L	0.0002	***

Cognitive Load	Diagram Understanding				Diagram Complexity			
	r	ES	p	SIG	r	ES	p	SIG
All Diagrams	−0.338	M	0.044	**	−0.081	S	0.640	
Bad Layout	−0.452	L	0.060	*	−0.313	M	0.207	
Good Layout	−0.197	S	0.434		0.152	S	0.548	

Table 2. Pearson's product-moment correlation between diagram size and modeler performance, measured as mean and variance of objective performance (correct answers, i.e., score), different subjective assessments, and cognitive load measures. In each cell, the first number is Pearsons's r indicating the size of the the correlation, the letter S/M/L classifies the effect size, the next number is the p-value, and the stars indicate its significance level.

Validity of Method & Results

- **Online survey vs. Paper**
  - Online surveys may reach a larger audience, but often achieve poor completion (~20%), high noise, and little control of participants.
  - Also, paper is more realistic, as domain experts and decision makers will usually be faced with printed model reports rather than modeling tools.
- **Questionnaires vs. Eye tracking**
  - Maletic et al. have used eye-tracking in UML CD comprehension studies to validate questionnaire-based results yielding similar observations.
  - Even modern eye tracking equipment imposes substantial difficulty and effort, thus severely restricting the number of participants.
- **Cognitive load measures**
  - Subjective assessments have been shown to be as reliable as physiological indicators such as skin conductivity, heart rate, pupillary dilatation.
- **Internal Validity**
  - Low p-values control Type I errors.
  - High n controls Type II errors.
- **External Validity**
  - Case studies and models are realistic, but not real.
  - Number of models might be too small for general conclusions.
- **Construct Validity**
  - Measuring cognitive load leads to different conclusions and a higher degree of construct validity than previous work.
- **Conclusion Validity**
  - Consistent observations over multiple measures, measurements, models, tasks, and populations.
  - Consistent results over a series of 7 experiments.
  - No independent replication yet.

Validity of Results

Participants

	male	female	all	completion rate (core questions)
novices	40	3	43	80.0 %
experts	30	4	34	84.4 %
all	70	7	77	81.9 %

Participants

EXPERIMENT	MALE	FEMALE	ALL	COMPLETION RATE CORE QUESTIONS
D (BENG)	29	3	33	75.1%
E (MSC)	29	5	34	82.6%
F (ELITE)	10	1	11	90.1%
ALL	68	10	78	80.5%

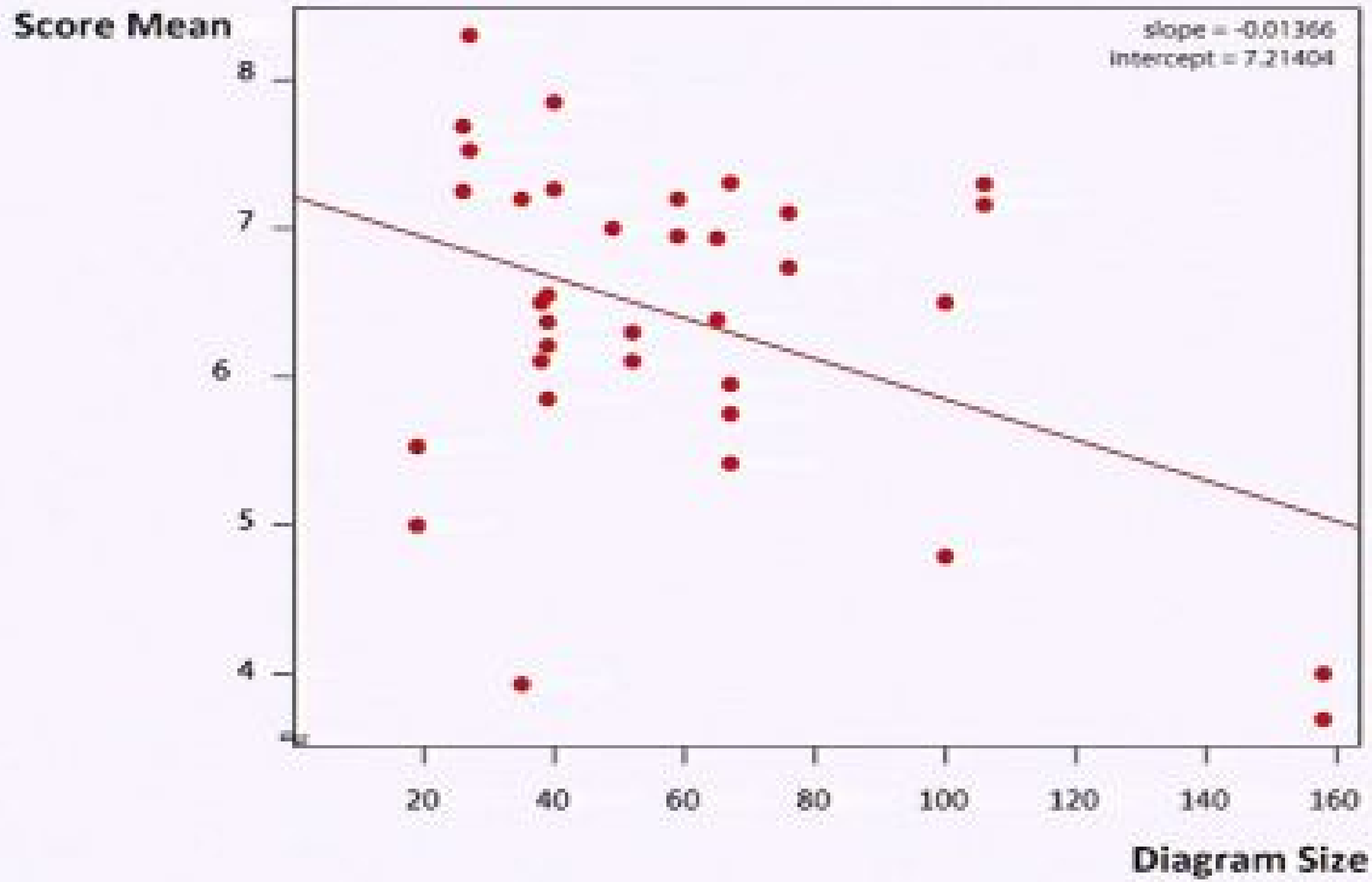
Significance

HYPOTHESIS	P-VALUE	SIGNIFICANCE
$H_{0,1}$: same user performance for good/bad layouts wrt.		
...correct answers	0.003	**
...wrong answers	0.002	**
...time per answer	0.061	.
...time per correct answer	< 0.001	***
$H_{0,2}$:same user assessment of good/bad layouts wrt.		
...layout quality	$< 10^{-15}$	***
...diagram clarity	$< 10^{-15}$	***
$H_{0,3}$ same performance for good/bad layouts by experts/novices wrt.		
...correct answers	< 0.0001	***
...wrong answers	< 0.0001	***
$H_{0,4}$: novices benefit more than experts from good layouts		
...correct answers	0.39	-
...wrong answers	0.24	-

Significance

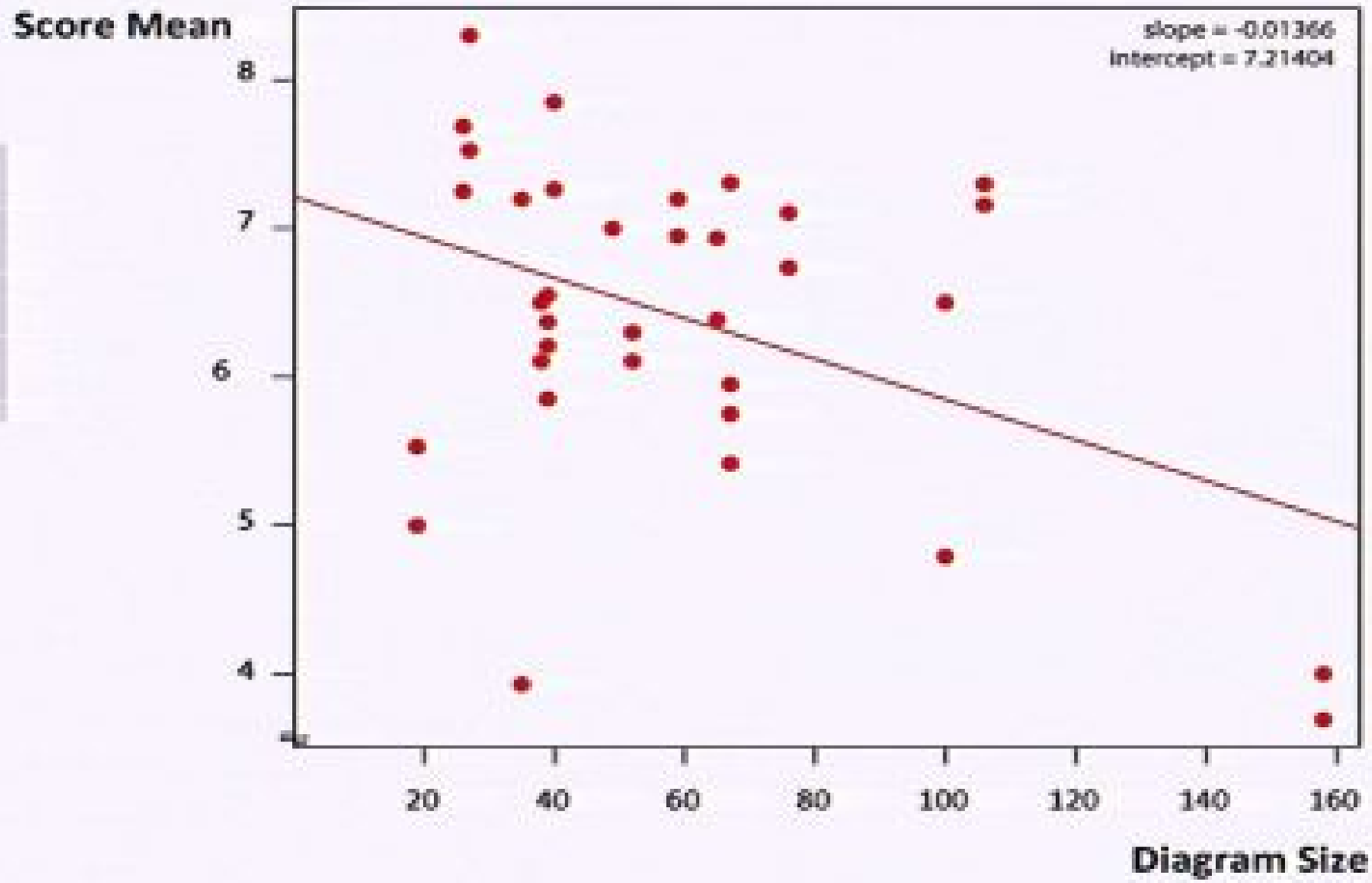
HYPOTHESIS	TEST	P-VALUE	SIGNIFICANCE
$H_{0,1}$: same user performance for good/bad layouts wrt.			
...correct answers	Wilcoxon	0.009	**
...time per answer	Wilcoxon	0.4	-
...time per correct answer	Wilcoxon	0.7	-
$H_{0,2}$:same user assessment of good/bad layouts wrt.			
...layout quality	Wilcoxon	$< 10^{-15}$	***
...diagram clarity	Wilcoxon	$< 10^{-9}$	***
$H_{0,3}$ same variability for user performance for good/bad layouts wrt.			
...correct answers	F	0.005	**
...time per correct answer	F	0.03	*
$H_{0,4}$ same variability for user assessment of good/bad layouts wrt.			
...layout quality	F	$< 10^{-9}$	***
...diagram clarity	F	0.009	**
$H_{0,5}$: same benefit from good layouts for different populations			
...D vs. E	Wilcoxon	0.003	**
...E vs. F	Wilcoxon	0.0013	**
...D vs. F	Wilcoxon	$< 10^{-4}$	***

Impact => Optimal Diagram Size



Impact => Optimal Diagram Size

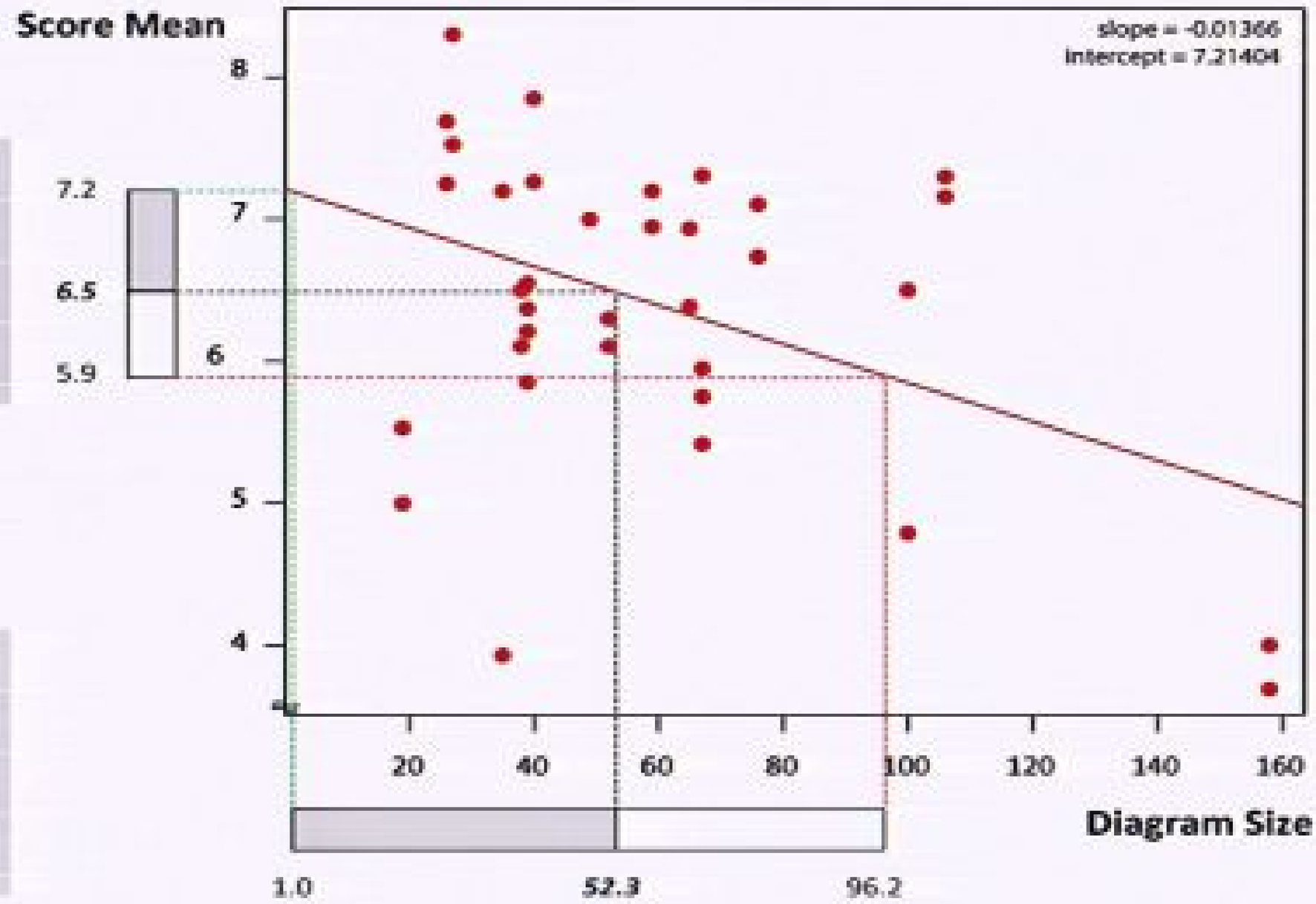
Score		Q1	Median	Q3
Layouts	Good	6.2	6.9	7.2
	All	5.9	**6.5**	7.2
	Bad	5.5	6.4	7.2



Impact => Optimal Diagram Size

Score		Q1	Median	Q3
Layouts	Good	6.2	6.9	7.2
	All	5.9	**6.5**	7.2
	Bad	5.5	6.4	7.2

Size		Q1	Median	Q3
Layouts	Bad	1.0	23.0	74.2
	All	1.0	***52.3***	96.2
	Good	1.0	59.6	125.5



Modeling Practices in Industry

- **There are many textbooks on modeling and model based software development, but much of what they describe does not resonate with my own practical experience.**
  - Examples are always tiny and tidy.
  - In reality, you don't use always use all models.

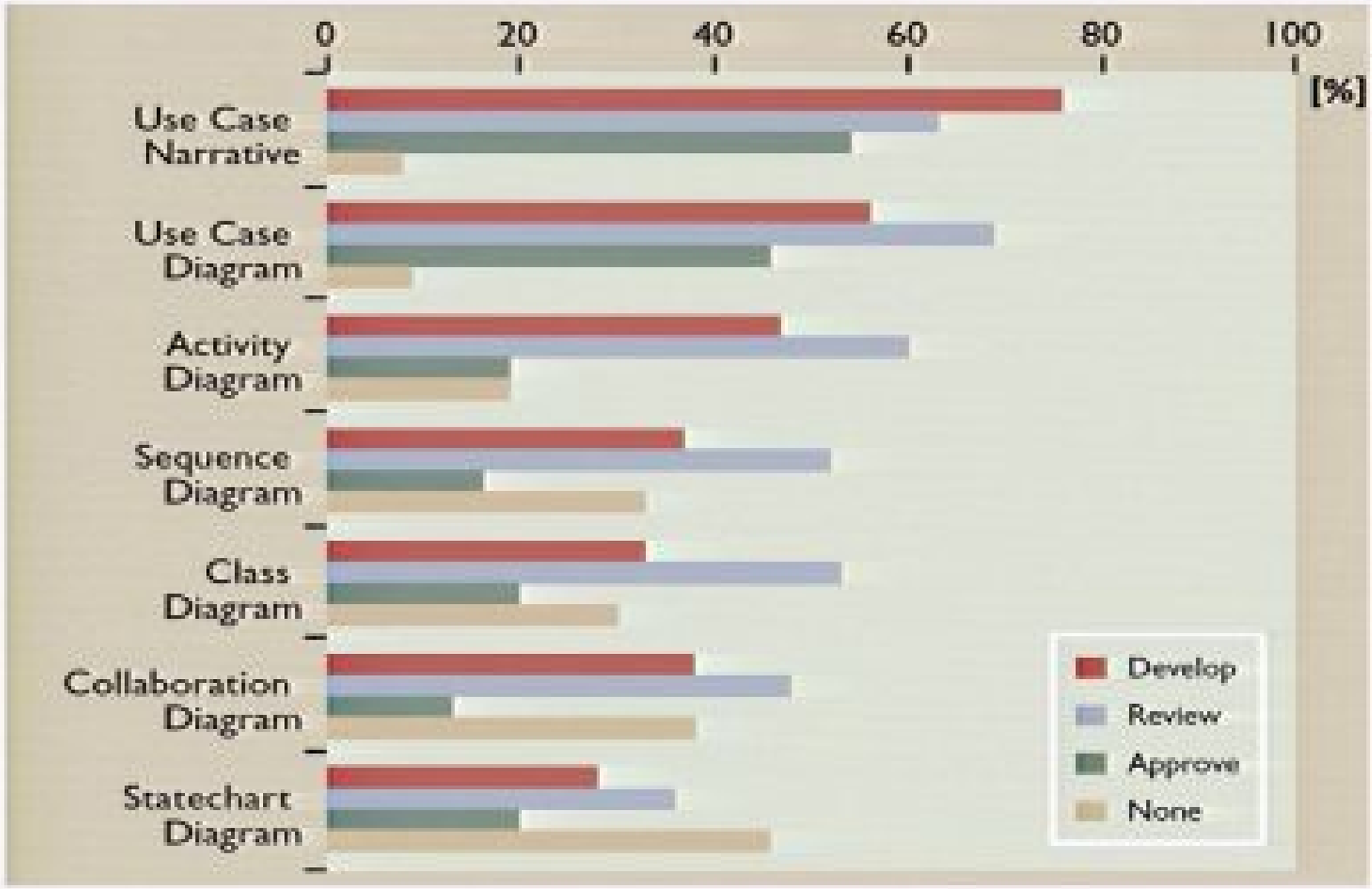
- **In a series of studies, Dobing and Parsons tried to find out which UML diagram types are actually used for what purposes and by whom, in industry.**

- **I am currently conducting a series of interviews, with very senior modelers from industry and investigate their modeling practices:**
  - what exactly are they doing
  - why are they doing it
  - what effects do they observe

- **Observing student modelers, I study influence factors on model usage.**
  - Modeling medium
  - Group composition
  - Constraints

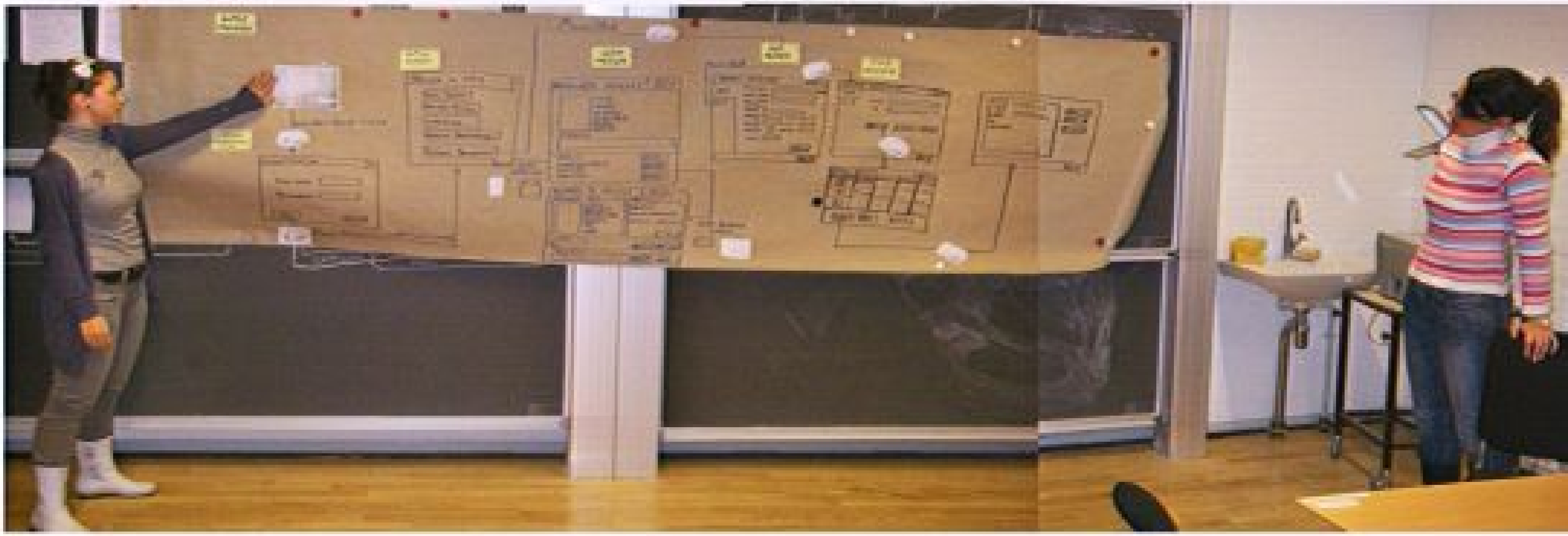
UML Diagrams used in the Analysis Phase

- **Dobing and Parsons studied which UML diagram types are actually used for what purposes and by whom, in industry.**



[Dobing, Parsons: How UML is used. CACM, May 2006 (49) 5, pp. 109—113]


Cancel
OK



Role of Empirical Research in Software Engineering

Early Stages
Mathematics and Electrical Engineering dominate Computer Science. "Software Engineering" is a contradiction in terms.

Contemporary SE
SE is scientifically accepted, but lacks industrial impact. Most work is conceptual, maybe implemented, but rarely validated empirically.

Future of SE
Practical work without empirical validation will be scientifically unacceptable. Industrial impact will grow due to solid evidence to support our claims.

1968
beginning of „Software Engineering"

1982
1st CHI Conference
HCI pioneers empirical techniques in CS

SIGCHI

1996
J.ESE is founded
4 issues annually

1997
1st EASE Conference

EASE
Evaluation and Assessment in SE

2007
1st ESEM Conference

2010
1st RESER Workshop

RESER
Replication in Empirical SE Research

2011
J.ESE increases frequency
impact factor= 1.227

x 6

Role of Empirical Research in Software Engineering

Early Stages
Mathematics and Electrical Engineering dominate Computer Science. "Software Engineering" is a contradiction in terms.

Contemporary SE
SE is scientifically accepted, but lacks industrial impact. Most work is conceptual, maybe implemented, but rarely validated empirically.

Future of SE
Practical work without empirical validation will be scientifically unacceptable. Industrial impact will grow due to solid evidence to support our claims.

1968
beginning of „Software Engineering"

1982
1st CHI Conference
HCI pioneers empirical techniques in CS

1996
J.ESE is founded
4 issues annually

1997
1st EASE Conference

2007
1st ESEM Conference

2010
1st RESER Workshop

2011
J.ESE increases frequency
impact factor= 1.227

x 6

MODELS Call for Papers

One of three categories of research papers
"Papers evaluating existing problem cases or scientifically validating proposed solutions through, for example, empirical studies, experiments, case studies, simulations, formal analyses, and mathematical proofs. [...] The research method must be sound and appropriate."

VL/HCC Call for Papers

"Research papers are expected to support their claims with appropriate evidence. [...] However, not all claims necessarily need to be supported with empirical evidence or studies with people. [...] Moreover, there are many alternatives to empirical evidence, [...] We encourage authors to think carefully about what claims their submission makes and what evidence would support them."

The Future of Sw. Development

Who do you want to rely on, heroes or scientists?

Be skeptical

There is no silver bullet.

Be patient

Empirical research is slow, but it produces real progress.

Be open

Allow us in: let us work with you.