Experiences from Empirical Studies at Simula Research Laboratory

Professor Dag Sjøberg

Simula Research Laboratory/

University of Oslo

Plan for the talk

1. Simula Research Laboratory
2. Motivation and goals for the SE work at Simula
3. Examples of technology transfer
4. Studies conducted at Simula
   - Experiments
   - Case studies
5. Sampling and recruitment of subjects
6. Supporting tools
7. Theory building
8. Data sharing
9. Challenges of empirical SE research

History of Simula Research Laboratory

1991:	Decision to close the airport at Fornebu, Oslo
1991-1997:	Political debate concerning possible use of the old airport
1999:	Funding approved for a research institute at Fornebu
2000:	The Parliament decides that IT-Fornebu shall develop a Knowledge Park at the old airport
2000:	Three research groups selected on basis of applications from 17 Norwegian university groups
2001:	Simula established
2004:	First Evaluation
2009:	Second Evaluation

Simula Research Laboratory

- 100 employees
- Shareholding company (Norwegian state: 80 %, Sintef and Norwegian computing centre: 20 %)
- Research departments:
  - *Networks and Distributed Systems*
  - *Scientific Computing*
  - *Software Engineering*
- Simula Innovation

Software Engineering Department at Simula

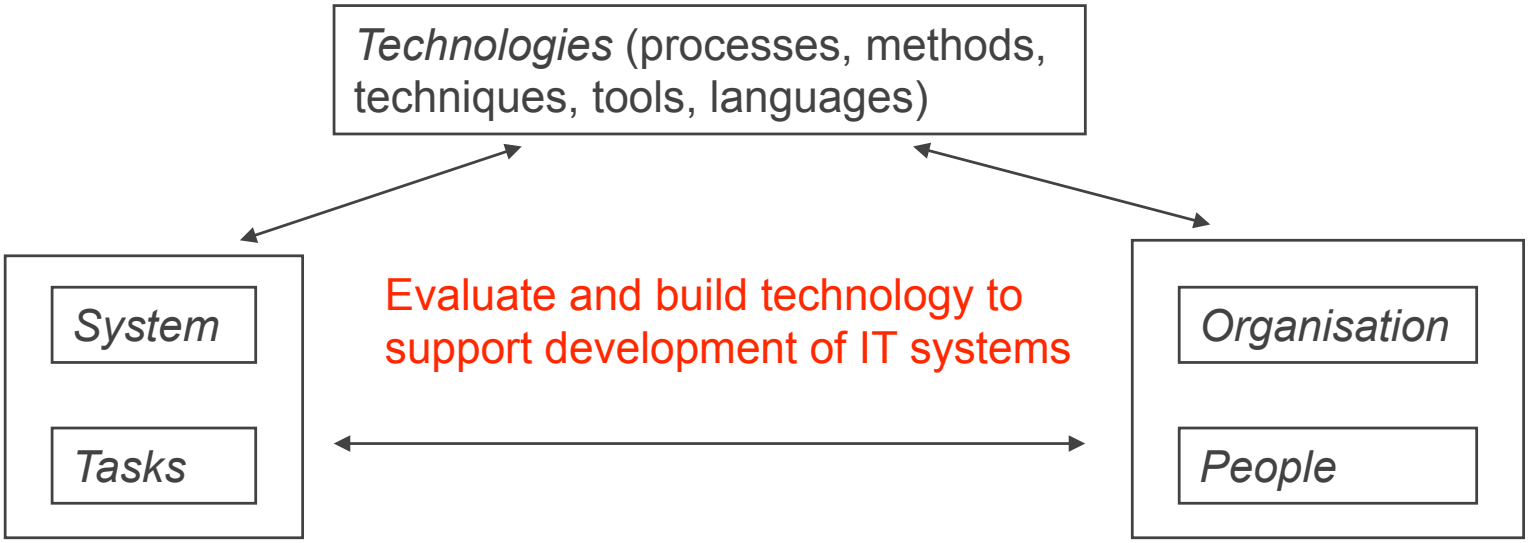
- 1999: “Industrial Systems Development” research group at University of Oslo
- 2001: SE Department of Simula
- Developing a research group almost from scratch combined with the resources available at Simula at that time, as well as strong management and focused research, created a unique opportunity
- The department is No. 3 (among 1361 institutions worldwide) in a ranking published in *Journal of Systems and Software*

Goal of Simula

- - to be an international leader in understanding and assessing the impact of SE technologies (processes, methods, techniques, languages and tools for building and maintaining software) on the human, organisational, and technological dimensions of systems development
- - and to
  - improve the state of the art of empirical research methods,
  - develop guidelines and tools for efficient collection and distribution of high quality empirical data, and
  - develop guidelines for theory building in software engineering

Motivation for SE research

The motivation for conducting research in software engineering is to support the private and public software industry in developing higher quality systems with improved timeliness in a more cost-effective and predictable way



- There may be hundreds of alternative technologies: How should the industry (and others who build software) judge what technologies are useful when?
- Many achievements have been made in the empirical SE community, but we are still far from generally being able to answer this question.

Two kinds of SE research

SE research is about

1. *developing* new, or *modifying* existing, technologies to support software development, or
2. *evaluating* and *comparing* the effect of using such technology in the often very complex interaction of
   - individuals
   - teams
   - projects and organisations
   - various types of tasks and software systems

Historically, activity (1) has been emphasised, but to make SE more scientific, much more effort is needed on activity (2).

Scientific Evaluation of SE Technology

- **Today: Mostly based on anecdotal evidence, personal opinion, arbitrary tests, etc.**
- **Sciences that study real-world phenomena use empirical methods by necessity, which involve systematic observation and experimenting, rather than deductive logic or mathematics**

Scientific challenges addressed at Simula

- Within the specific areas (effort estimation, testing, model-driven development, maintenance, process improvement) Simula aims to
  - to *quantify* and *understand* the effect of using various process models, methods, techniques and tools in various industrial situations, that is, provide a cost-benefit analysis over variation in software developers, teams, projects and organisations, and various types of activities and software system.
- In the areas where we have a fair understanding of the effect, we also propose new or modified technologies (e.g., in the area of software effort estimation)
- To have impact on practice/industry, software engineering research must involve the industry: "The industry is our lab"

How do we collaborate with industry?

Type of collaboration		N comp.	N pers.
Empirical studies with professional practitioners as participants	Experiments (from 15 minutes to 8 weeks)	262	2730
	Case studies (a few days to 3 years)	8	25
	Action research/case studies (up to 5 years)	9	84
	Interviews (typically 1 hour)	18	63
Research collaboration	Grant applications	18	36
	Achieved grants	26	74
	Co-authoring scientific papers	8	24
Technology transfer	Giving courses	4	101
	Giving seminars	33	1189
Teaching SE courses	Guest lectures from industry at courses given by the dep.	10	31
Acquiring consultancy work	E.g. to build infrastructure and organise studies	12	22
Total		**408 (326 unique)**	**4379**

Two examples of technology transfer

Case Study on the use of UML-based development in ABB within the SPIKE project

The company ABB joined the SPIKE project in 2003 because they wanted to increase productivity and quality in their product development projects through the use of UML-based development.

The project has resulted in, among others, the following publications:

B. C. D. Anda, K. Hansen, I. Gullesen, and H. K. Thorsen. Experiences from Using a UML-based Development Method in a Large Safety-Critical Project, Empirical Software Engineering 11(4):555-581, 2006.

N. E. Holt, B. C. D. Anda, K. Asskildt, L. C. L. Briand, J. Endresen, and S. Frøystein. Experiences with Precise State Modeling in an Industrial Safety Critical System, In: Critical Systems Development Using Modeling Lanuguages, CSDUML'06, ed. by Siv Hilde Houmb, Geri Georg, Robert France, Dorina C. Petriu, and Jan Jürjens, chap. 6, pp. 68-77, Springer, 9th edition ed. (ISBN: 0809-1021), 2006.

For ABB the cooperation with Simula has resulted in requirements documents that are significantly improved.

ABB is a global company and a small increase in the precision of their deliveries means savings of several million NOK per year.

Improving Estimation Practices at the Norwegian Directorate of Taxes

Over the last years several small experiments have been conducted in industry on how various factors, for example, expectations of clients and developers and amount of information in the requirements documents affect software estimates.

The joint results of these experiments have been presented to software developers at the Norwegian Directorate of taxes in a course over 4*2 hours.

The plan is that this shall improve the estimates of the "SL -project", which is a large project developing the new system for calculating taxes for individuals.

Controlled experiments

- One contribution of the empirical SE community is the conducting of experiments to evaluate and compare industrial SE technologies
- How do we convince practitioners and managers in industry that the results of controlled experiments are relevant to them?
- The applicability of the experimental results to industrial practices is in most cases hampered by the experiments' lack of *realism* and *scale* regarding subjects, tasks, systems and environments, that is, the challenge of achieving external validity

State of the art in SE experimentation

*Sjøberg *et al.*, A survey of controlled experiments in software engineering, IEEE Transactions on Softw. Engineering 31(9) (2005), pp. 733–753.

		Articles reporting controlled experiments	
Journal/ Conference	Total no. of articles investigated	N	Row %
EMSE	124	22	17.7
ISESE	20	3	15.0
METRICS	177	10	5.6
JSS	886	24	2.7
TSE	687	17	2.5
ICSE	520	12	2.3
IST	745	8	1.1
SME	186	2	1.1
IEEE SW	532	4	0.8
TOSEM	125	1	0.8
IEEE Comp	780	0	0
SP&E	671	0	0
All	**5453**	**103**	**1.9**

Definition of experiment

- "Controlled experiment in software engineering (operational definition): A randomized experiment or a quasi-experiment in which individuals or teams (the experimental units) conduct one or more software engineering tasks for the sake of comparing different populations, processes, methods, techniques, languages, or tools (the treatments)"

- Excluded are correlation studies, studies that are solely based on calculations on existing data (e.g., from data mining), and evaluations of simulated teams based on data for individuals. The last category falls outside our operational definition because the units are constructed after the run of the experiment.

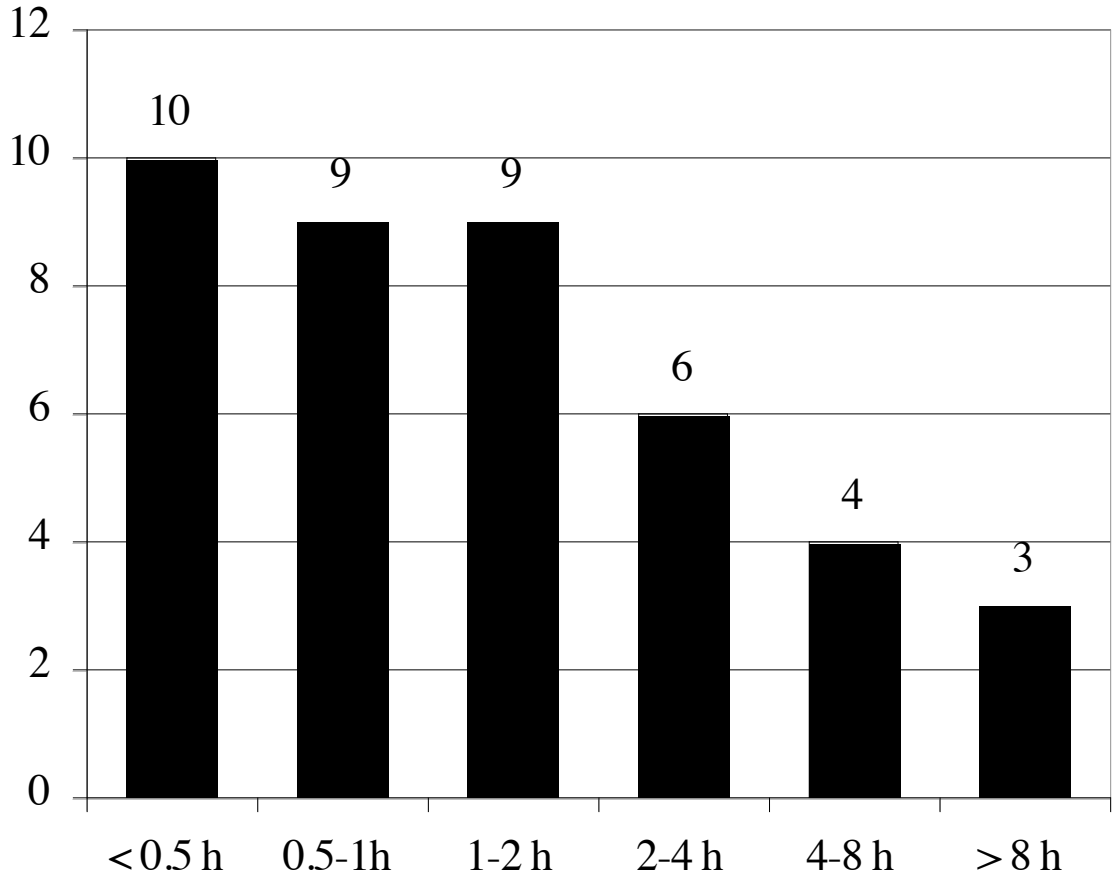
Subjects

Subject Category	Reported Subject Types	N	%
Undergraduates	Undergraduates , Bachelors , Third and fourth -year students, Last-year students, Honors and Majors .	2969	54.1
Graduates	Graduate students , S tudents following graduate courses or Master's programs , MSc and PhD students .	594	10.8
Students, type unknown	Students in computer science, S tudents .	1203	21.9
Professionals	Developers, P ractitioners, Software engineers, Analysts , Domain experts, Business managers , Facilitators , Professionals.	517	9.4
Sci entists	Professors, Post-doctorates , Staff members of educational institutions .	74	1.3
Unknown		131	2.3
Total		5488	100

Realism (representativeness) of tasks, systems and environments

- A grand challenge in SE experimentation is how we generalise from the specific tasks, systems and environments of SE experiments
- Not aware of suitable taxonomy or classification of these aspects for SE
- Nevertheless, development tasks in industry usually take longer and are often more complex than is the case in most experiments

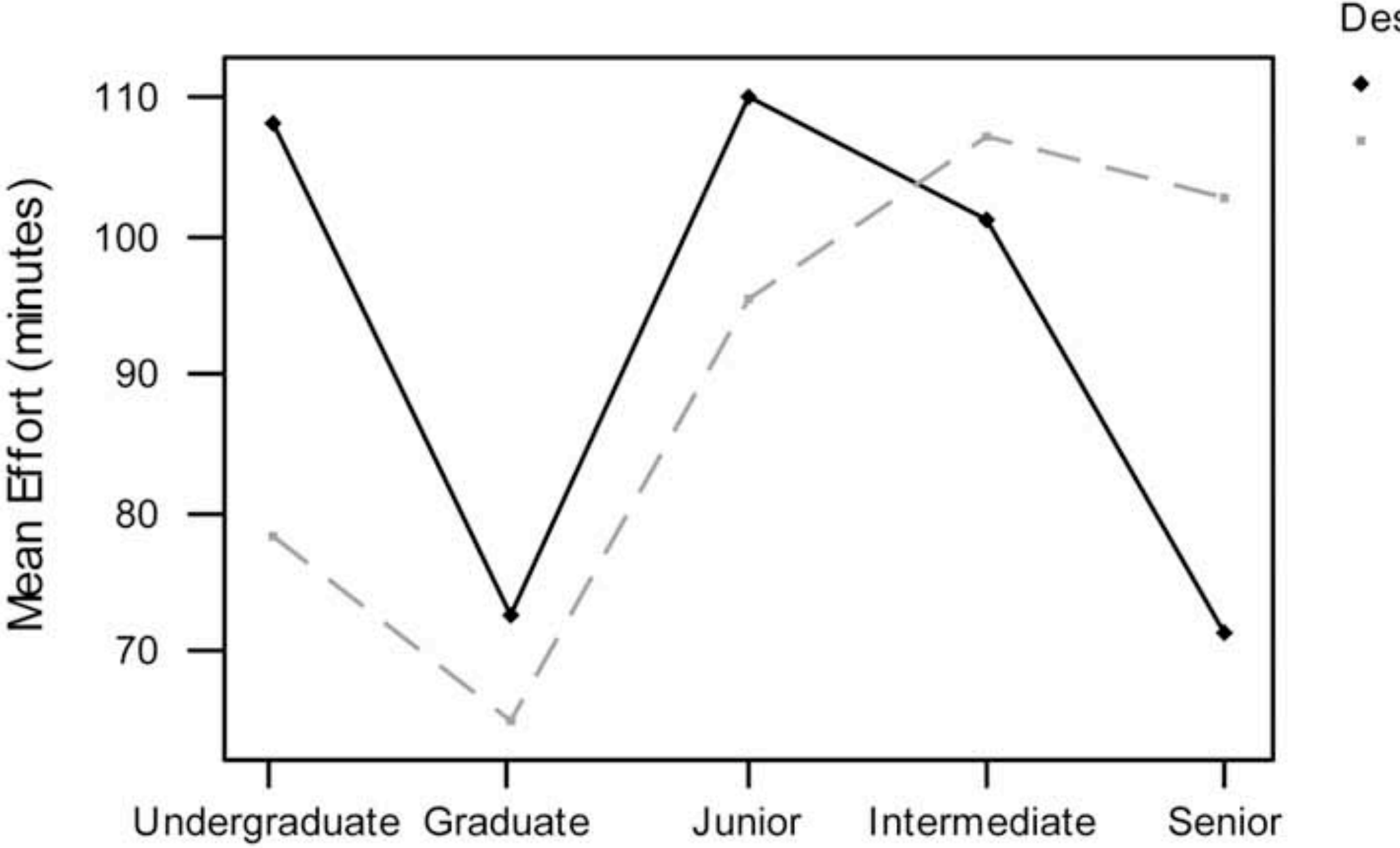
Duration of experiments with time measurements



Why is scale important?

- Easier to obtain a representative sample of the target population.
  - One of 113 experiments reported sampling from a well defined target population
- Many aspects of the complexity of software engineering only manifest themselves in controlled experiments if the experiments involve a sufficiently large number of subjects and tasks, for example, differences among subgroups of subjects

Is a helicopter better than a bike?


Des
110
100
90
80
70
Mean Effort (minutes)
Undergraduate
Graduate
Junior
Intermediate
Senior

Another example, (quasi) experiment on pair programming

295 *junior*, *intermediate* and *senior* professional Java consultants from 29 companies were paid to participate (one work day)

99 individuals (conducted in 2001/2002)

98 pairs (conducted in 2004/2005)

- Norway: 41
- Sweden: 28
- UK: 29

The pairs and individuals performed the same Java change tasks on either:

- a ”*simple*” system (centralised style) or
- a ”*complex*” system (delegated style)

We measured duration (elapsed time), effort (cost) and correctness

Why that many subjects? Power analysis

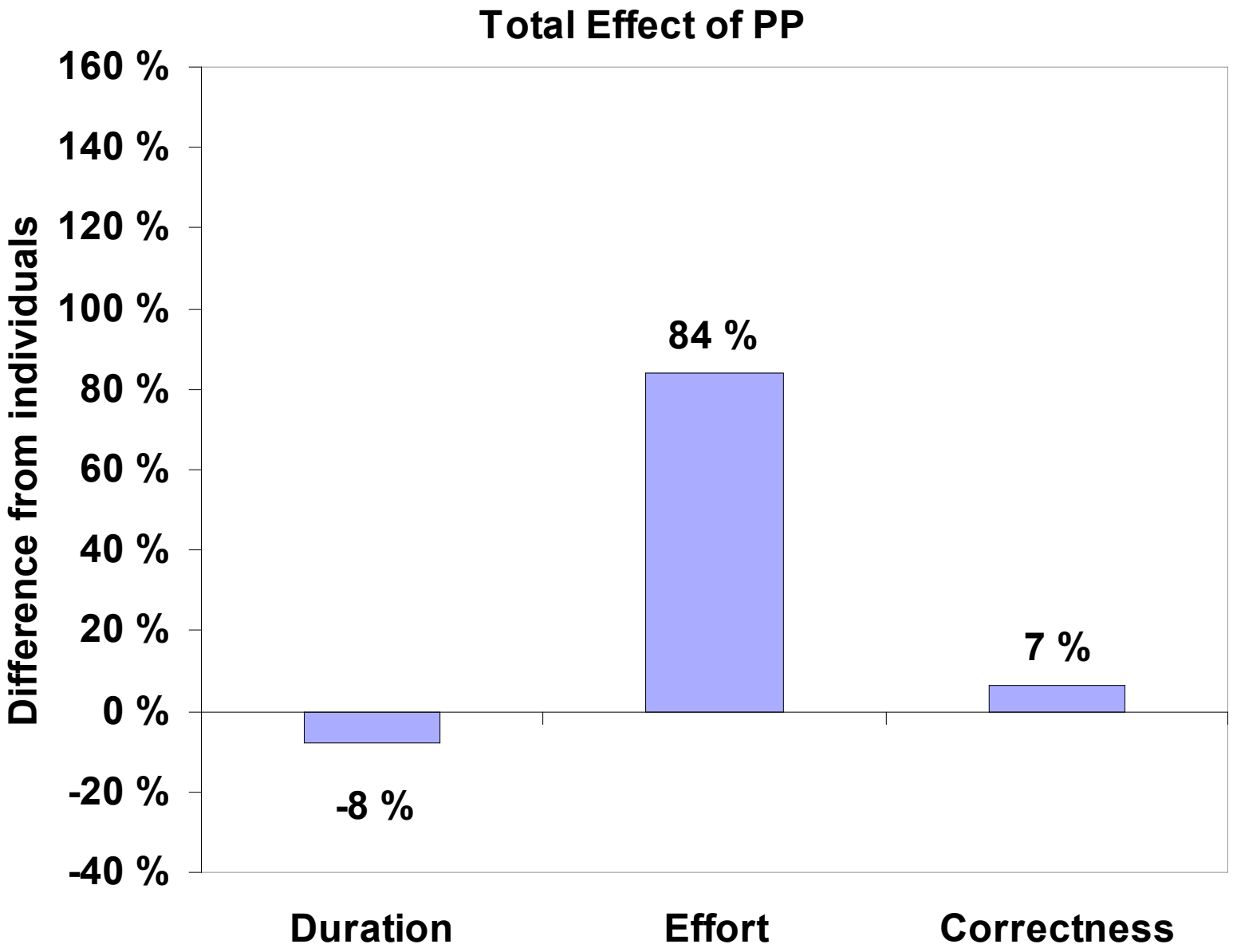
Research question:

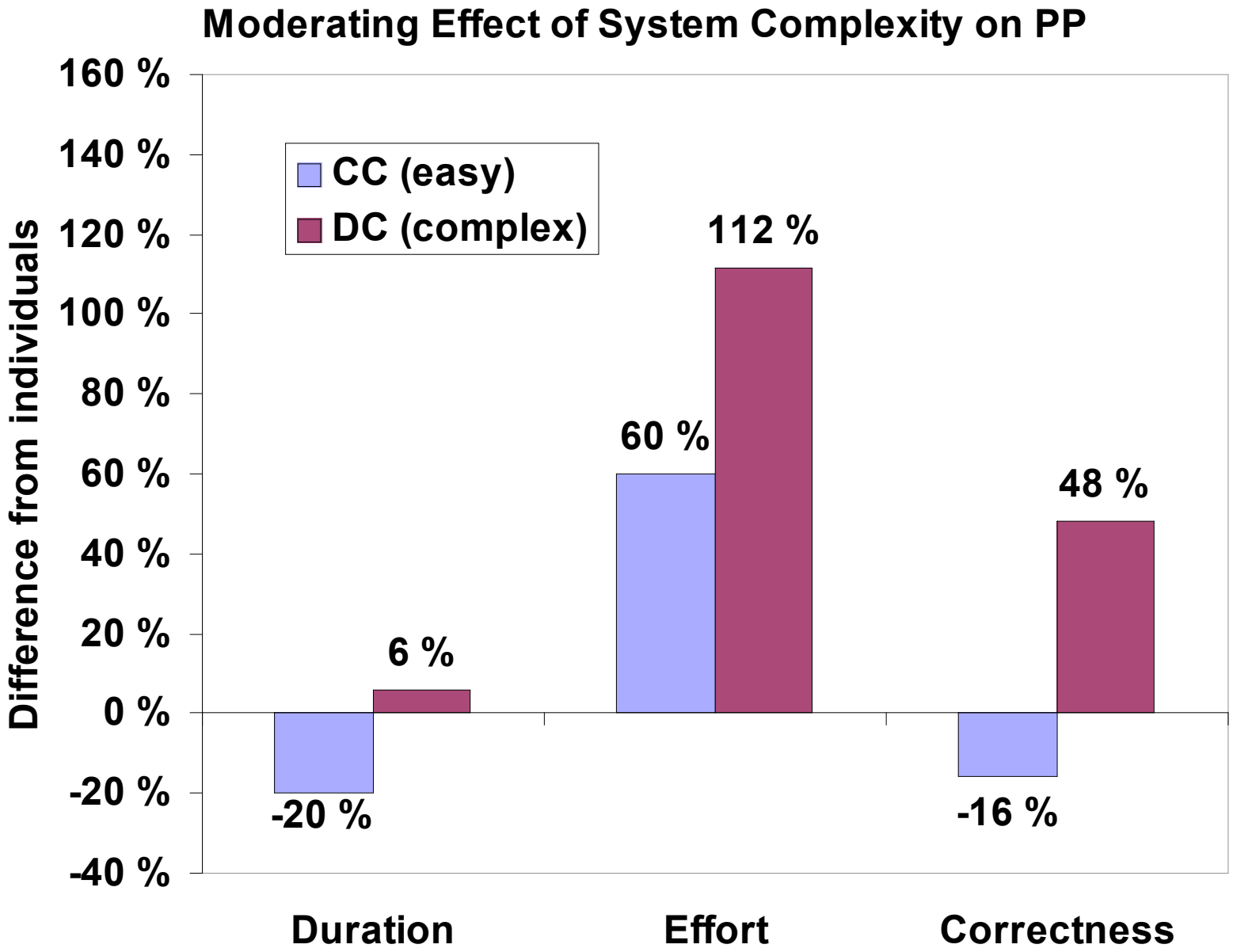
What is the effect regarding duration, effort and correctness of pair programming for various levels of system complexity and programmer expertise when performing change tasks?

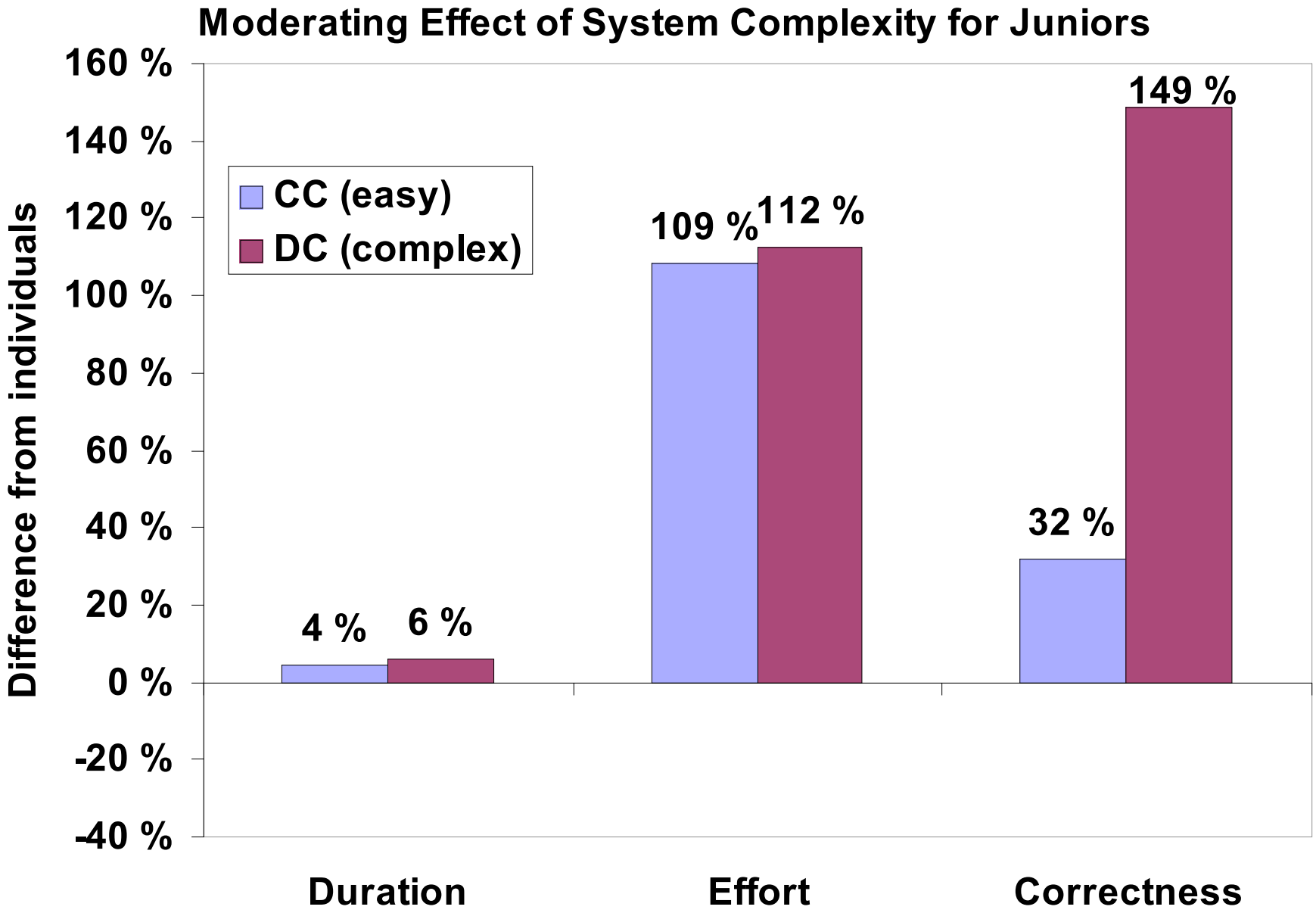
2x2x3 fixed-effect analysis of covariance:
pair programming (two levels), control style (two levels) and expertise (three levels), resulting in twelve levels/groups

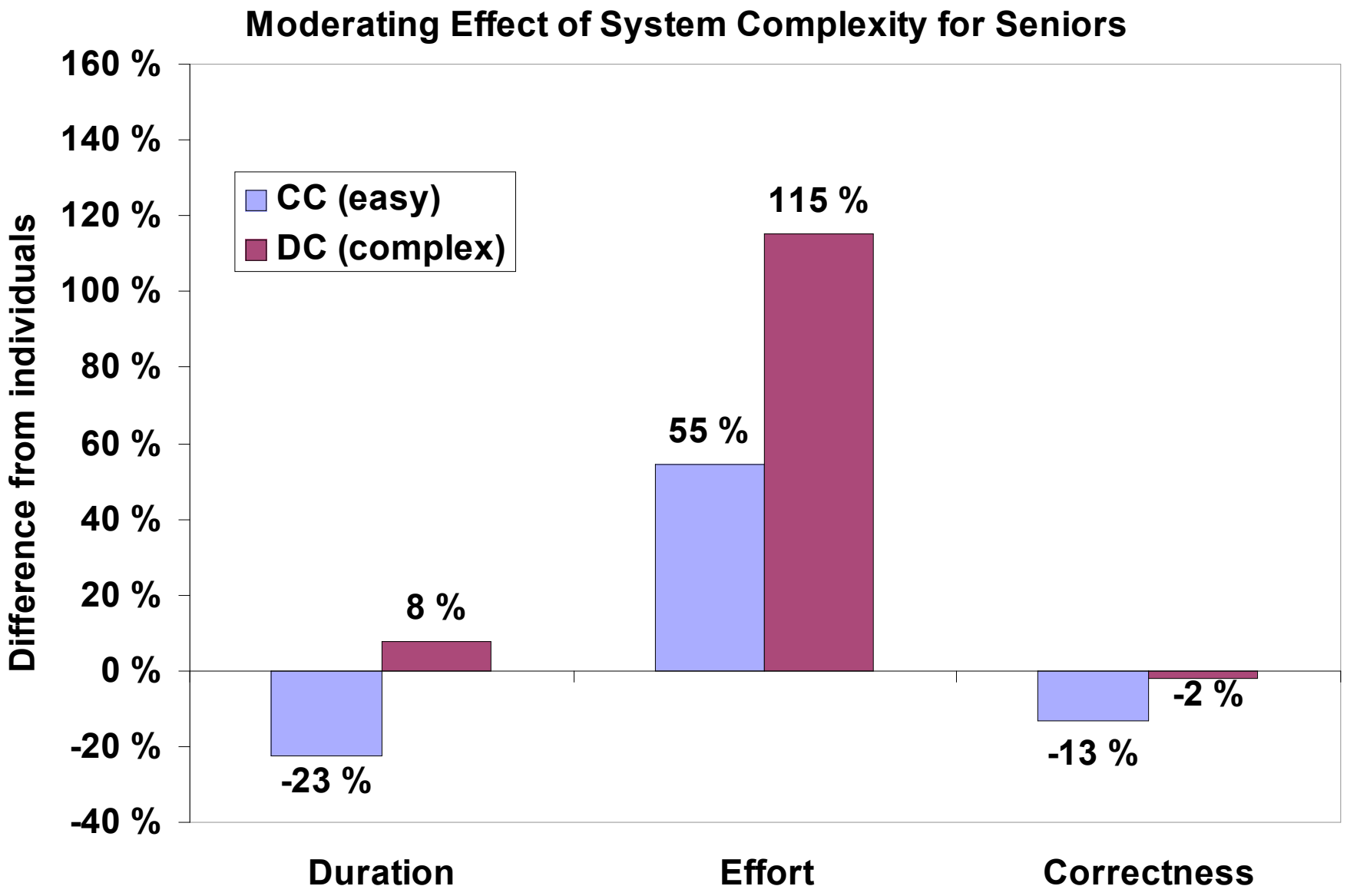
***N* = 170 (85 individuals and 85 pairs)**

N = 14 in each of the 12 groups


Total Effect of PP
Difference from individuals
160 %
140 %
120 %
100 %
80 %
60 %
40 %
20 %
0 %
-20 %
-40 %
84 %
7 %
-8 %
Duration
Effort
Correctness


Moderating Effect of System Complexity on PP
Difference from individuals
160 %
140 %
120 %
100 %
80 %
60 %
40 %
20 %
0 %
-20 %
-40 %
CC (easy)
DC (complex)
-20 %
6 %
60 %
112 %
-16 %
48 %
Duration
Effort
Correctness


Moderating Effect of System Complexity for Juniors
Difference from individuals
160 %
140 %
120 %
100 %
80 %
60 %
40 %
20 %
0 %
-20 %
-40 %
CC (easy)
DC (complex)
4 %
6 %
109 %
112 %
32 %
149 %
Duration
Effort
Correctness


Moderating Effect of System Complexity for Seniors
Difference from individuals
160 %
140 %
120 %
100 %
80 %
60 %
40 %
20 %
0 %
-20 %
-40 %
CC (easy)
DC (complex)
-23 %
8 %
55 %
115 %
-13 %
-2 %
Duration
Effort
Correctness

The effect of PP "depends on"

Programmer expertise	Task complexity	Use PP?	Comments
Junior	Easy	Yes	Provided that increased quality is the main goal
	Complex	Yes	Provided that increased quality is the main goal
Intermediate	Easy	No	
	Complex	Yes	Provided that increased quality is the main goal
Senior	Easy	No	
	Complex	No*	

* Unless you are sure that the task is too complex to be solved satisfactorily even by solo seniors

The performance of the various categories may depend on their relevant education, work experience, the actual task and system, development technology, etc.

In the survey of 113 experiments, 7 involved both students and professionals. Only 3 measured difference in performance: partly no difference, partly professionals better.

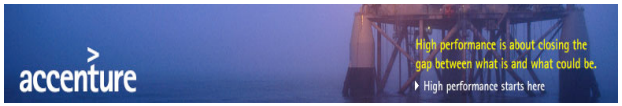
How to run large experiments with professionals?

"practitioners are understandably skeptical of results acquired from a study of 18-year-old college freshmen."

"finding 100 developers willing to participate in such an experiment is neither cheap nor easy. … But even if a researcher has the money, where do they find that many programmers?"

[W. Harrison, "Skinner Wasn't a Software Engineer", Editorial, *IEEE Software,* May/June, 2005]

Industry relationship since 2001:
326 companies and public institutions



Examples of experiments at Simula

99 consultants from 8 companies

- one-day experiment that compared two different object-oriented control styles

295 consultants from 29 companies in Norway, Sweden and the UK

- one-day experiment that tested the effect of pair programming

39 consultants from 11 companies

- Three-day experiment on design patterns

20 programmers from 13 companies

- worked individually from one to two weeks in an experiment on UML

35 companies presented bids for a web-based system that we needed

- 4 were selected to actually build the system independently of each other.
- The teams (2-3 developers from each company) spent from 7 to 25 person-weeks each

30 companies from 11 countries in Europe and Asia presented their bids.

- 4 companies built the system
- each spent from 10 to 20 person-weeks

Reward mechanisms in SE experiments

	Experiment		Participant	
Reward	**N**	**%**	**N**	**%**
Grade	10	8.8	732	13.3
Extra credits	9	8.0	660	12.0
Payment *	3	2.7	121	2.2
Other rewards	1	0.9	24	0.4
No reward	16	14.4	458	8.3
Unknown	74	65.5	3493	64.6
Total	**113**	**100**	**5488**	**100**

*Only students

Incentives for industry to collaborate

Type of collaboration	Type of study	Incentive	N comp.	N pers.
Empirical studies	Experiments	Simula (and partly research council) pay industry	158	1094
		Simula give seminar/increased knowledge	125	1636
	Case studies	Simula pay industry	8	25
	Action research/ case studies	Simula offer expertise, Research council pay Simula's time (40%), industry spend own time (60% of total costs) to improve business processes. 1 comp. funds 1 PhD	8	83
Taking part in research & innovation projects		Improving own business processes/ increased knowledge	26	74
Teaching SE courses		Promoting company and individual?	10	31
Acquiring consultancy work		Simula pay	12	22

Hiring consultants

- The experiments listed above cost between €50,000 and €230,000
- We paid the companies ordinary consultancy fees for individuals or fixed price for a whole project, like any other ordinary customer.
  - The companies have routines for defining (small) projects with local project management, resource allocation, budgeting, invoicing, providing satisfactory equipment, etc.
- Difficult to find subjects employed in an in-house software development company because the management will typically prioritize the next release of their product

How do we get the money?

- At Simula, the research and administrative leader is the same person
- Relatively few constraints on how we spend the money as long as we can envisage a good research outcome
- Decided to use 25% of budget for experiments, mainly at the expense of employing a larger number of researchers

Apply for money to conduct experiments

- Finding the money to fund comprehensive experiments is a matter of politics. How many apply to funding bodies for money to pay for professionals to take part in experiments?
- In research grants applications, we budget for money for positions, equipment and travel; why not include money for experiments?
- Compared with large projects in other disciplines, e.g., physics and medicine, we are talking about a relatively small amount of money

Subject recruitment from companies

- Finding companies: Internet or specialized databases of vendors, associations, partners and interest groups indicate whether their profile matches the defined target population
- "What's in it for us?" Some organizations, commercial value. For others, new knowledge and skills would be more attractive.
- From a given organization we will require some number of individuals, with some profile, at some time and location and for some duration.
  - Participation more attractive if company can utilize temporary demand dips
  - Using an internet-based experiment tool (SESE), we can offer flexibility regarding location
  - Flexibility re number of participants from each organization

[H. C. Benestad, E. Arisholm and D. Sjøberg. How to Recruit Professionals as Subjects in Software Engineering Experiments, In: IRIS (Information Systems Research in Scandinavia), 6-9 Aug., Kristiansand, Norway, 2005]

Communication strategy

Making contacts:

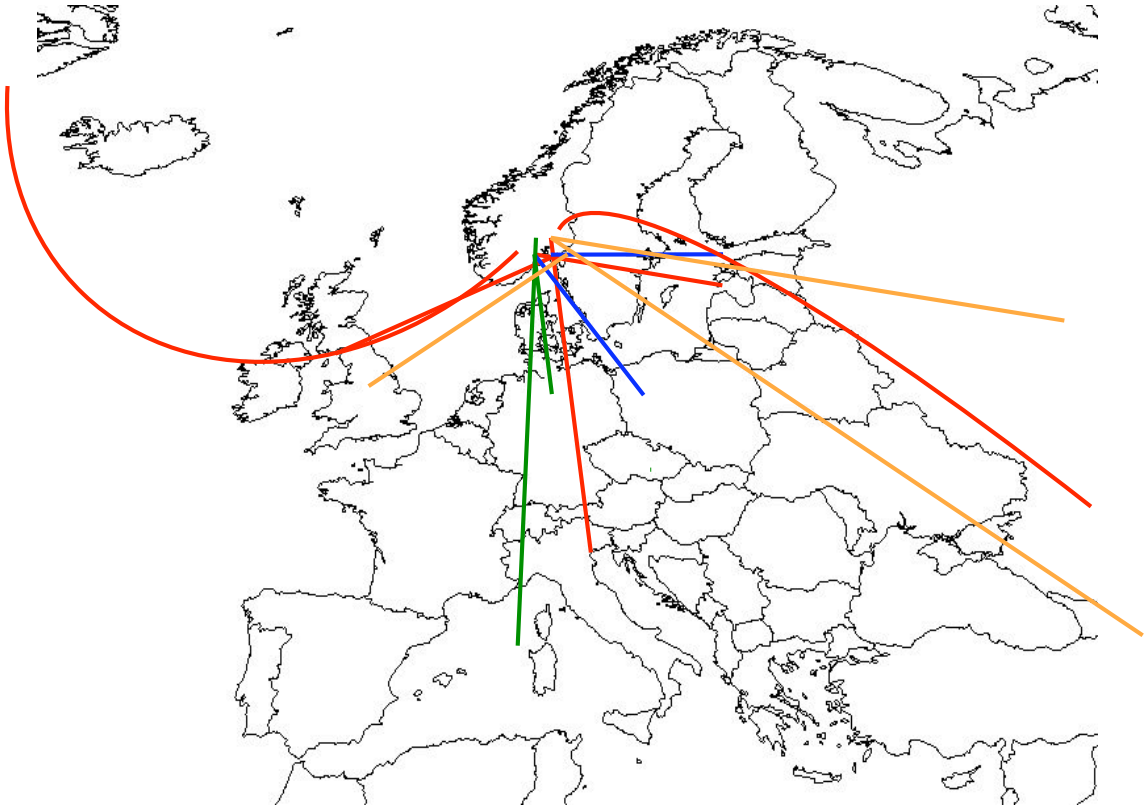
- Who makes the first enquiry?
  - an enquiry from a research director may have greater effect one from a research assistant
- Who is the first enquiry made to?
  - It works well to make enquiries to the higher levels of company hierarchies, identifying people that control resources of interest to the project, and that normally handle external relations.
  - Using the switchboard and asking questions like “Who manages java resources in your company” has been successful. For smaller companies the CEO. If the contact point belongs to a low level of the company, there is a danger of selecting from a specific sub-culture, or the person may not have necessary incentive or power to attract potential participants.
- What medium is used for the first enquiry?
- email can be perceived as less intrusive and can be well planned. For people with good communicative qualities, phone calls can be effective

Examples of prepared materials that can be are:

Prepared materials

- An email template for the first contact
- A contract template
- General information on the experiment, with requirements to the participants
- Information/checklist to the coordinator
- Information/checklist to the participants

Empirical studies with professionals – a global activity

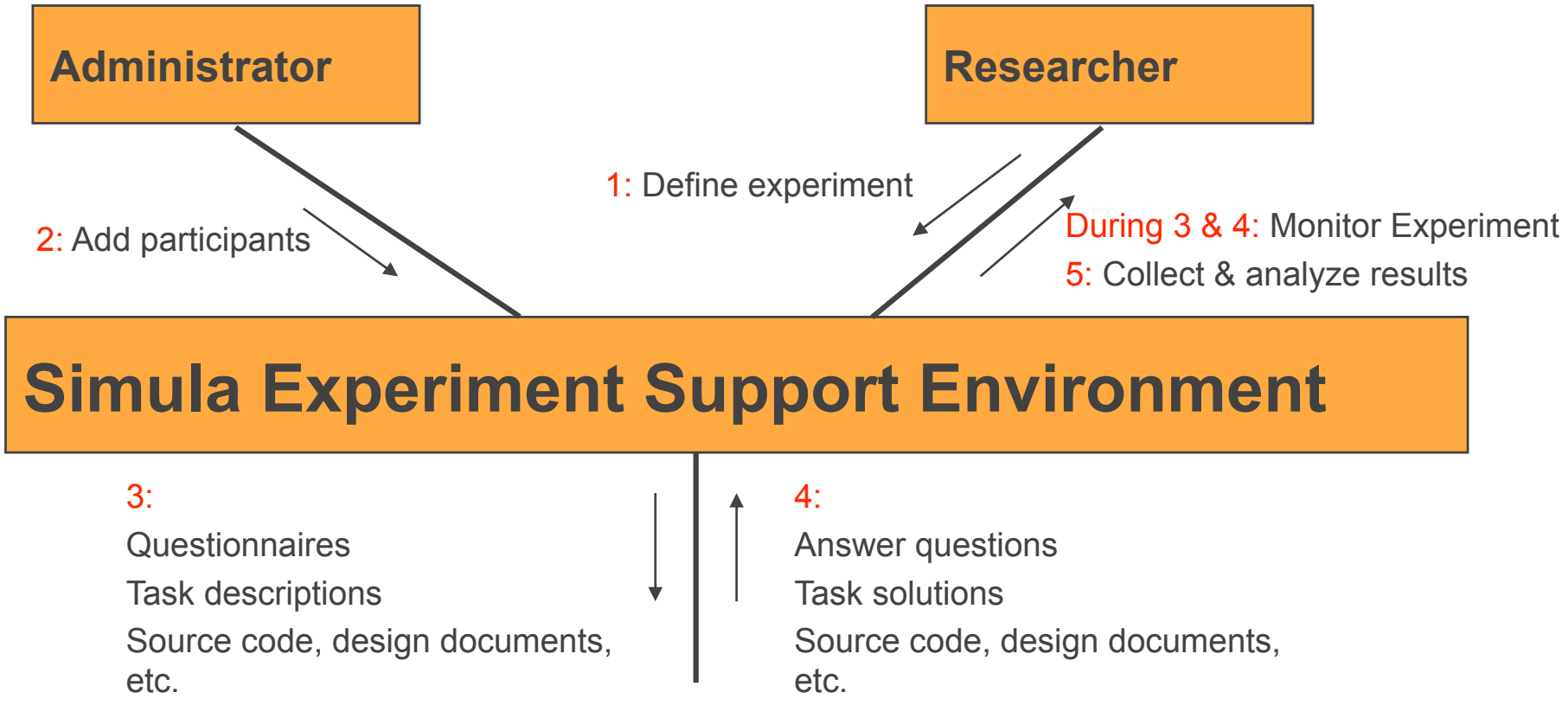


Country	Companies	People
Norway	216	3427
India	18	110
Russia	17	45
Sweden	14	161
Ukraine	7	20
Pakistan	7	14
UK	5	60
Romania	5	57
Nepal	4	101
Belarus	4	45
Bulgaria	4	8
Denmark	3	79
Vietnam	3	77
Germany	2	80
Ukraina	2	80
Poland	2	23
Czech Rep.	2	21
Moldovia	2	15
Italy	1	11
Finland	1	10
Lithuania	1	10
China	1	2
Phillipines	1	2
Serbia	1	2
Slovakia	1	2
Thailand	1	2
Canada	1	1
27	**326**	**4465**

The logistics of controlled experiments is work intensive and error prone

- Personal information and background information of subjects must be collected
- General information and specific task documents must be printed and distributed
- Solution documents must be collected

Web-based tool support (SESE)



SESE is also used for surveys

Key functionality of SESE

- real-time monitoring of the experiment
- flexibility of defining new kinds of questions and measurement scales
- automatic recovery of experiment sessions
- automatic backup of experimental data
- multi-platform support for downloading experimental materials and uploading task solutions

SESE is built on top of a commercial human resource management system, and is partly being developed by an external company

[E. Arisholm, D. I. Sjøberg, G. J. Carelius and Y. Lindsjørn. A Web-based Support Environment for Software Engineering Experiments, Nordic Journal of Computing 9(4):231-247, 2002.]

Practical organisation of large experiments

- Ask for a local project manager of the company who selects subjects according to the specification of the researchers, ensures that the subjects actually turn up, ensures that the necessary tools are installed on the PCs, and carries out all other logistics, accounting, etc.
- Motivate the experiment up-front: inform the subjects about the purpose of the experiment (at a general level) and the procedure (when to take lunch or breaks, that phone calls and other interruptions should be avoided, etc.).
- Ensure that the subjects do not talk with one another in breaks, lunch, etc.
- Ensure the subjects that the information about their performance is kept confidential (both within company and outside).
- Ensure the company that its general performance is kept confidential.
- Monitor the experiment, that is, be visible and accessible for questions.
- Give all subjects a small training exercise to ensure that the PC and tool environment are working properly.
- Ensure the company and subjects that they will be informed about the results of the experiment.
- Provide a proper experiment support environment that is used to set up and monitor the experiment, and collect and manage the experimental data.

Professional Project Manager

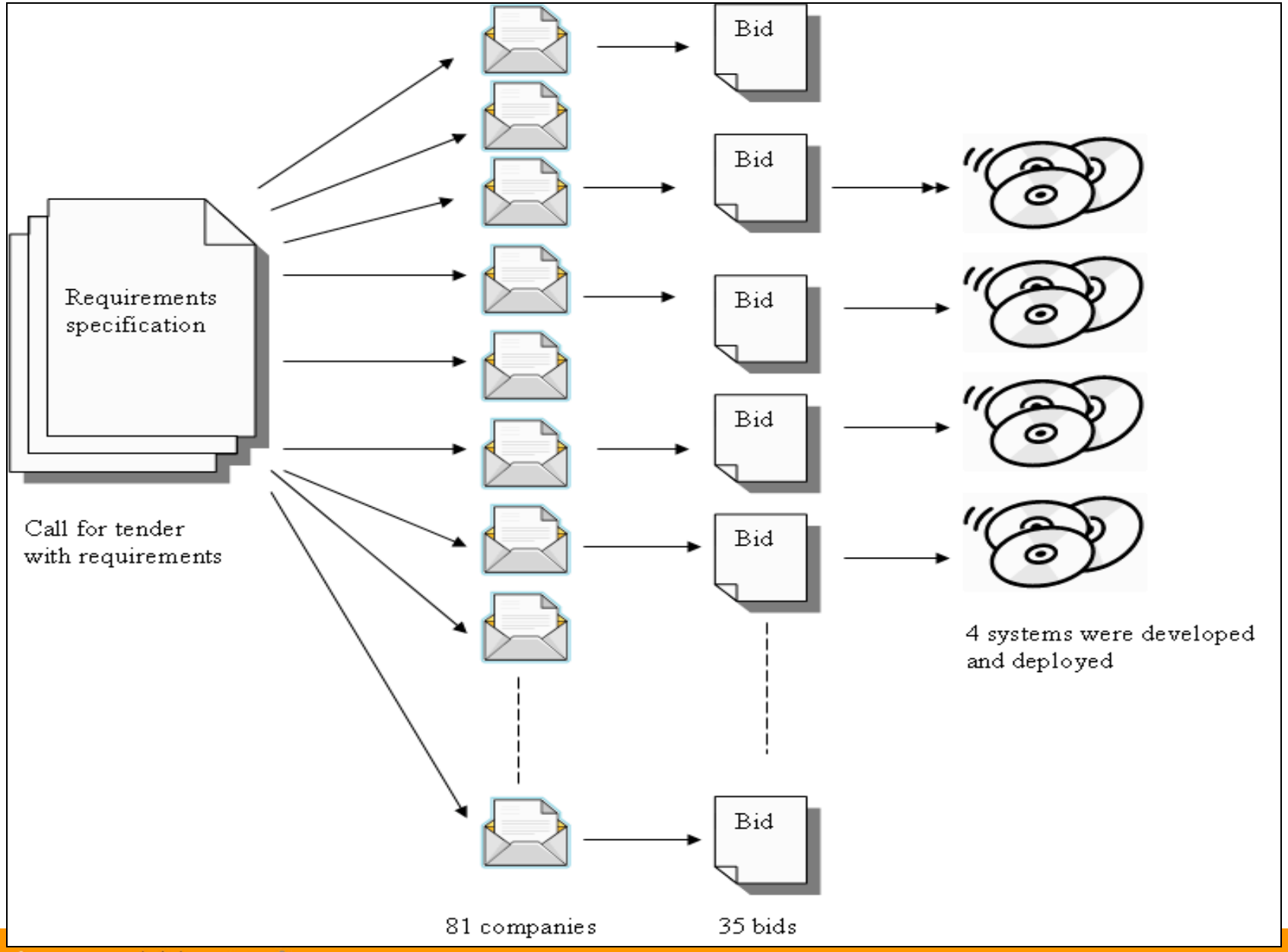
Simula Scientific Advisory Board, January 2003:

> **"The recent strategy of paying for professional services, e.g. the implementation of SESE, the analysis of experiments, arranging subjects for experiments, etc. is clearly a good policy and should be continued. However, this does involve a great deal of planning, supervision, interaction and administration. … Taking on public leadership exacerbates these workloads. Dag Sjøberg is in greatest danger from such distracting workloads. It will therefore be worthwhile appointing a project and operations manager, who is sufficiently senior to take the initiative in conducting this work, and to arrange that he or she develops a team that can undertake the planning, external arrangements, conduct, etc. of experiments and can also take a major responsibility for arranging that data is properly curated."**

June 2003:

> **We successfully employed as Knowledge and Project Manager, who used to be the Development Manager for one of the largest case tool vendors in Scandinavia.**

A multiple-case study



DES part 1: Full realism exp. on bidding

The bidding process consisted of two separate phases:

In pre-study phase, 17 of the 35 bidding companies indicated price based on an incomplete description of user requirements

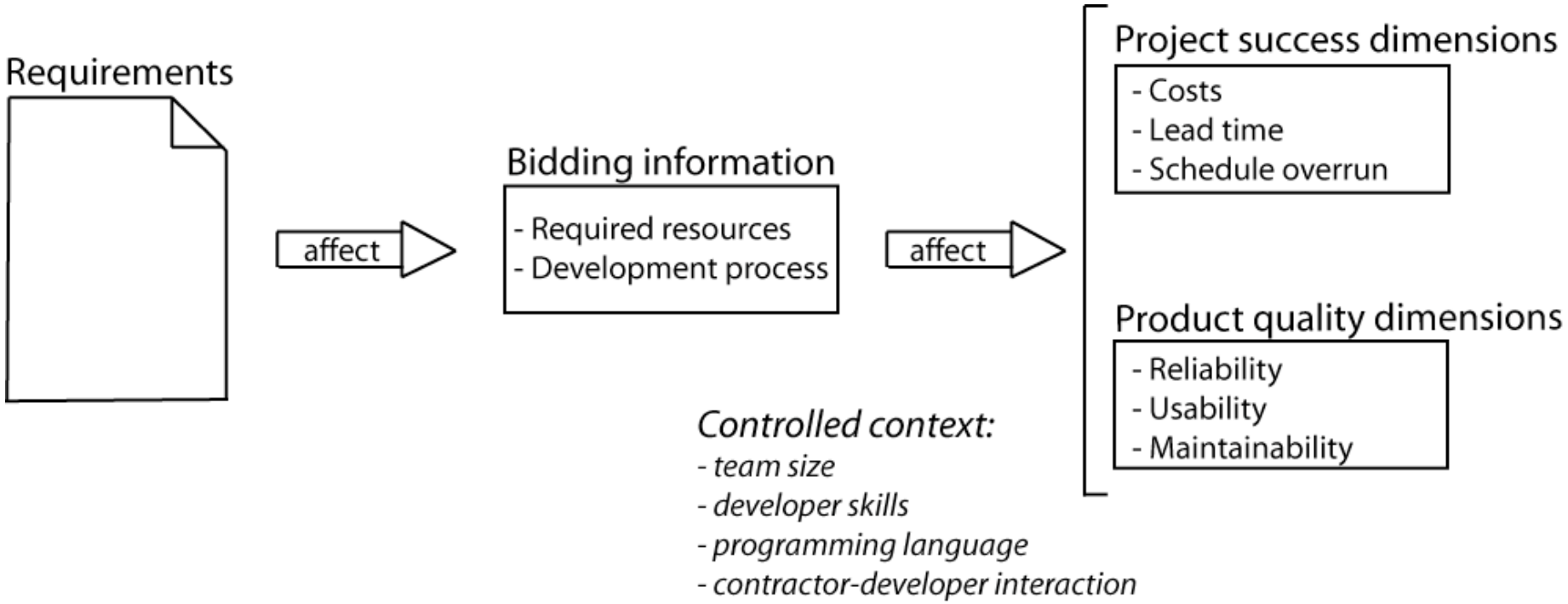
In the bidding phase, all 35 companies provided bids based on a more complete requirement specification with substantially more functionality than the system indicated in the pre-study phase

The 17 companies involved in the pre-study phase presented bids 70% higher than the bids of the other companies.

Preliminary theory:

1) Software clients tend to achieve better price/uncertainty relationships, i.e., better prices, when the requirement uncertainty perceived by the bidders is low.
2) Software clients should not request early price indications based on limited and uncertain information when the final bids can be based on more complete and reliable information

DES phase 2: Multiple-instance case study with much control



Studies variability/reproducibility

- The *firm price*, *planned schedule*, and *planned development process*, had, respectively, “low”, “low”, and “medium” reproducibility.
- The contractor’s *costs*, *actual lead time*, and *schedule overrun* had, respectively, “medium”, “high”, and “low” reproducibility
- *Reliability, usability*, and *maintainability* of the delivered products had, respectively, “low”, “high”, and “low” reproducibility
- Variability for predictable reasons is also included in the notion of reproducibility. The observed outcome matched expectations, formulated on the basis of SE folklore, to some extent. Nevertheless, achieving more reproducibility in SE remains a great challenge for SE research, education, and industry.

B. C. D. Anda, D. Sjøberg, and A. Mockus. Variability and Reproducibility in Software Engineering: A Study of four Companies that Developed the same System, Accepted for IEEE Transactions on Software Engineering, 2008.

Data source	Description
Access logs	The logs from Simula's web server were collected during the two years the systems were operational.
Bids	The companies' offered firm price was included in all the bids. The time schedule of the project, the planned development process, and the analysis and design of the product were included in many of the bids.
Contractor time sheets	Simula's contractor team recorded the time they spent on the project.
CVS	At the completion of the development and testing, the researcher team received complete CVS bases from the projects.
E-mail	All the e-mail communication between Simula's project manager and the development projects was recorded.
Interviews	The projects' team members were interviewed weekly about their work on the project and about the possible effects of being the object of research. The interviews were semi-structured and based on an interview guide in which some questions were the same each week, while others varied depending on the status of their project.
Issue tracker	The companies registered their questions and needs for clarification in Bugzero; see http://www.websina.com/bugzero/. The Simula contractor team registered their responses. Later on, the Simula team registered defects (classified according to severity) that were detected in the acceptance tests.
Log of defects	This is the log of the defects found after the systems became operational.
Project documents	These are documents related to overall project management, such as time schedules, design descriptions, acceptance test logs, and technical documentation.
Running Systems	The four systems.
Snapshots	Snapshots of all documents (including code) were sent to the research team weekly.
Source code	The source code of the four systems

Looking forward: What are the main challenges of the empirical SE community?

- More empirical studies
- Higher quality studies
  - More relevant studies
  - More valid studies (construct, internal, external and statistical conclusion validity)
    - Identifying the context (moderator) variables of subjects and objects that may affect the results. These variables should then be used to characterise populations. Defining the scope of validity of our experiments is necessary to *compare* and *generalise the results*
- More focus on synthesizing evidence
- Theory building

Current SE research literature

- Percentage of articles that report empirical studies :
  - Tichy: 17%
  - Glass et al.: 14%
  - Sjøberg et al.: 12-17%
- Primary studies
  - Controlled experiments 1.9% (Sjøberg et al.)
  - (Personal opinion) Surveys 1.6% (Glass *et al*.)
  - Case studies 12% (Holt)
  - Action research 0% (Glass)
- Reviews and meta-analysis: 1-3% of papers
- Rough estimate: 180 studies a year

Need: ~2000 studies

- Assume there are 1000 research questions of high industrial importance that are meaningful to decide empirically, and
- assume that a research question requires at least 20 high quality studies, conducted over the last 10 years.
- This requires that we conduct at least 2000 high-quality empirical studies every year.

See more details in: D.I.K. Sjøberg, T. Dybå and M. Jørgensen. The Future of Empirical Methods in Software Engineering Research, In Future of Software Engineering (FOSE '07), edited by Briand L. and Wolf A., Minneapolis, US, 23-25 May 2007. IEEE-CS Press, pages 358-378, 2007.

More empirical studies

State of Practice	Target (2020-2025)
Relatively few empirical studies in SE research. Focus on developing new technology	Large number of studies covering all important fields of SE and using different empirical methods. Most research that leads to new or modified technology is subject to empirical evaluation
Empirical methods not part of industrial practice	Most large software development organizations conduct empirical studies as part of decisions making and process improvement

More relevant studies

State of Practice	Target (2020-2025)
Few results answer questions posed by industrial users, e.g., “Which method should we use in our context?” Current focus is on comparing mean values of technologies without a proper understanding of individual differences or the studied population	More focus on individualized results, individual differences, and better descriptions of populations and contexts; why, when and how is technology X is better than Y
Reference points for comparisons of technologies are frequently not stated, or not relevant	New technology is compared with relevant alternative technology used in the software industry
One may question the industrial relevance of many SE studies	More case studies and action research. Experiments should show more realism regarding subjects, technology, tasks, and software systems

More Valid Studies

Internal validity

The *internal validity* of an experiment is "the validity of inferences about whether observed co-variation between *A* (the presumed treatment) and *B* (the presumed outcome) reflects a causal relationship from *A* to *B* as those variables were manipulated or measured" [Shadish, 2002]. Changes in *B* may have alternative causes than the manipulation of *A*. An alternative cause for the outcome is a *confounding factor*.

More Valid Studies

Construct validity

- We need to measure something to understand it, but just as importantly, we need to understand something in order to measure it. What can be measured meaningfully in SE?
- For example: Quality = number of errors? What about functionality, usability, maintainability, etc. And what kind of errors, found where, found when? Compared with what?
- In general, low construct validity in SE studies, although little systematic investigation on this issue. Simula plans to carry out a systematic review in this area

More Valid Studies

External validity – Generalisation

The validity of inference about whether the cause-effect relationship holds over variation in:

- **Actors: individual, teams, project,organisation or industry**
- **Technology: process model, method, technique, tool or language**
- **Activities: plan, create, modify or analyze (a software system)**
- **Software systems: many dimensions, such as size, complexity, application domain, business/scientific/ student project or administrative/embedded/real time, etc.**

Dimensions of Generalization

	Statistical generalization	**Analytical generalization**
Individual studies	Statistical hypothesis testing	Generalization through theory or analogy
Collection of studies	Meta analysis	Research synthesis, aggregation of evidence, and theory

Generalization

State of Practice	Target (2020-2025)
The scope of validity of empirical studies is rarely defined explicitly	The scope is systematically and explicitly defined and reported
Statistics-based generalization is the dominant means of generalization	Studies include a diverse and reflected view on how to generalize, particularly through the use of theory

More Valid Studies

Statistical conclusion validity

The validity of inferences about the correlation (covariation) between treatment and outcome.

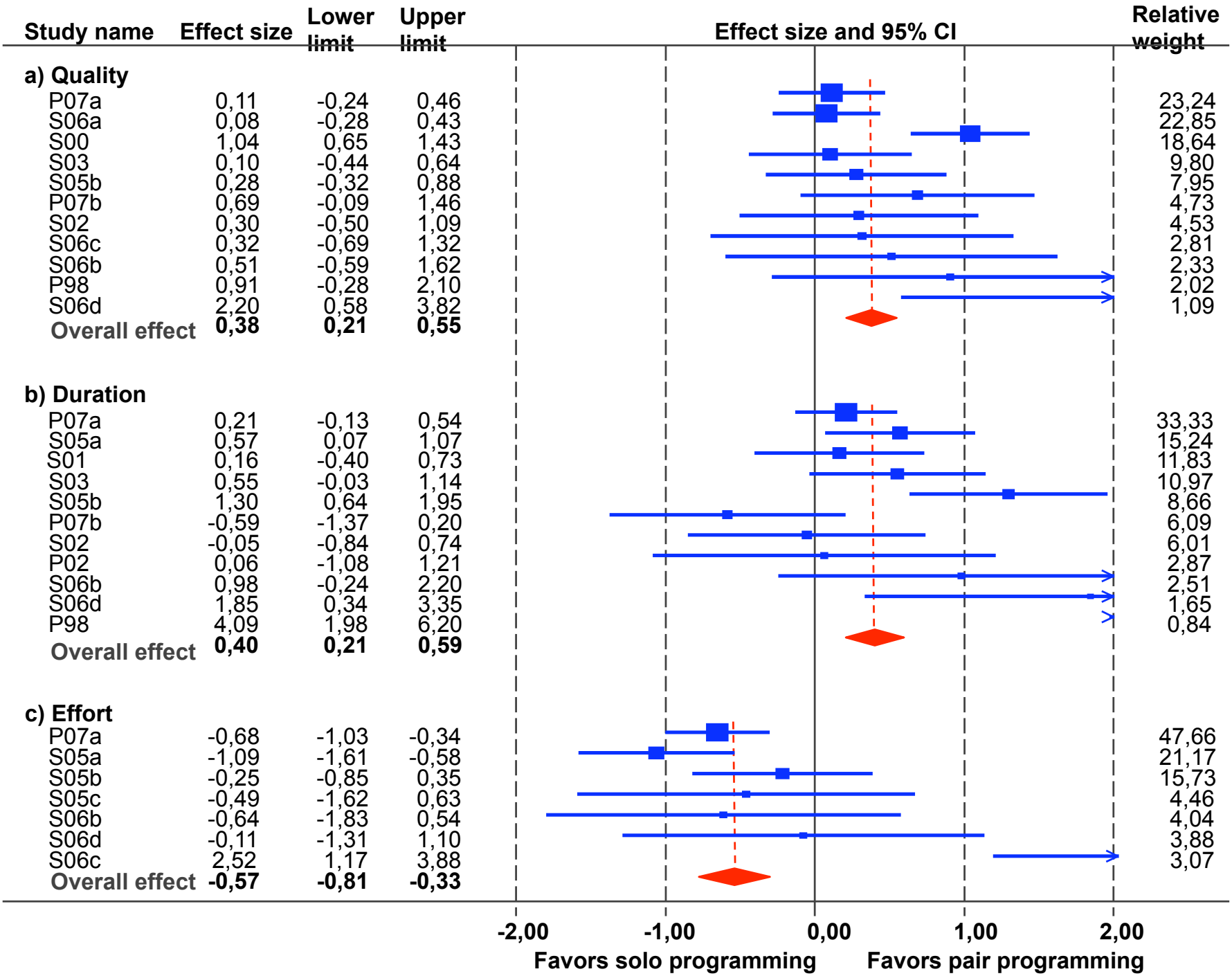
- **Statistical power is the probability that a statistical test will correctly reject the null hypothesis. A test without sufficient statistical power will not provide enough information to draw conclusions regarding the acceptance or rejection of the null hypothesis.**
- **An effect size quantifies the effects of an experimental treatment. Whereas *p*-values reveal whether a finding is *statistically* significant, effect size indicates *practical* significance, importance, or meaningfulness.**

Statistical Conclusion Validity

State of Practice	Target (2020-2025)
Stat. methods are used mechanically, with little focus on limitations and assumptions. Populations not defined, and for experiments, lack of power analysis and effect size estimation.	The use of statistical methods is mature. Populations are well defined, and power analysis and effect size estimation are conducted when appropriate.

Synthesizing Evidence

- *Primary*: collection and analysis of data
  - Experiments, surveys, case studies, action research, and others
- *Secondary:* research synthesis, summary, integration and combination of the findings of different primary research studies on a certain topic
  - Systematic reviews, meta-analysis


Study name
Effect size
Lower limit
Upper limit
Effect size and 95% CI
Relative weight
a) Quality
P07a 0,11 -0,24 0,46 23,24
S06a 0,08 -0,28 0,43 22,85
S00 1,04 0,65 1,43 18,64
S03 0,10 -0,44 0,64 9,80
S05b 0,28 -0,32 0,88 7,95
P07b 0,69 -0,09 1,46 4,73
S02 0,30 -0,50 1,09 4,53
S06c 0,32 -0,69 1,32 2,81
S06b 0,51 -0,59 1,62 2,33
P98 0,91 -0,28 2,10 2,02
S06d 2,20 0,58 3,82 1,09
Overall effect 0,38 0,21 0,55
b) Duration
P07a 0,21 -0,13 0,54 33,33
S05a 0,57 0,07 1,07 15,24
S01 0,16 -0,40 0,73 11,83
S03 0,55 -0,03 1,14 10,97
S05b 1,30 0,64 1,95 8,66
P07b -0,59 -1,37 0,20 6,09
S02 -0,05 -0,84 0,74 6,01
P02 0,06 -1,08 1,21 2,87
S06b 0,98 -0,24 2,20 2,51
S06d 1,85 0,34 3,35 1,65
P98 4,09 1,98 6,20 0,84
Overall effect 0,40 0,21 0,59
c) Effort
P07a -0,68 -1,03 -0,34 47,66
S05a -1,09 -1,61 -0,58 21,17
S05b -0,25 -0,85 0,35 15,73
S05c -0,49 -1,62 0,63 4,46
S06b -0,64 -1,83 0,54 4,04
S06d -0,11 -1,31 1,10 3,88
S06c 2,52 1,17 3,88 3,07
Overall effect -0,57 -0,81 -0,33
-2,00
-1,00
0,00
1,00
2,00
Favors solo programming
Favors pair programming

Systematic literature reviews

"The success of the Simula Research Laboratory in applying the principles of EBSE and performing high quality SLRs is supported by the strategy of constructing databases of primary studies related to specific topic areas and using those databases to address specific research questions. A database of cost estimation papers from over 70 journals [16] has been the basis of many of the detailed cost estimation studies authored or co-authored by Jørgensen and the database of 103 software experiments [36] has allowed researchers to assess a number of specific research trends in software experimentation."

[B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering – A systematic literature review, *Information and Software Technology*, Vol. 51, No. 1, pp. 7-15, January 2009]

The need for conceptual models/theories

Isolated hypotheses:

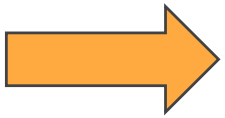
Technology (process, method, technique, tool language) *A* is better than technology *B*

Model/theory:

When and why is *A* better than *B*?
Depending on category of developers, tasks, systems, company culture and other environmental factors, *A* is better than *B*, because …

Use and sharing of theories

- **Of 5,453 articles, 103 report controlled experiments. 24 of those use a total of 39 theories to explain the cause-effect relationship under investigation**
- **Only 2 of the extracted theories are used in more than one article, and only 1 of these is used in articles by different authors**



little sharing of theories, even within topics

J. E. Hannay, D. I. K. Sjøberg, and T. Dybå. A Systematic Review of Theory Use in Software Engineering Experiments, IEEE Transactions on Software Engineering 33(2):87-107, 2007.

Theory building

State of Practice	Target (2020-2025)
Generally, little use of theories. The theories used mainly justify research questions and hypotheses; some explain results; very few test or modify theory	Most SE studies involve theories. Considering using, testing, modifying or formulating theory is part of any empirical work
Almost no SE-specific theories are proposed	Many SE theories are proposed and tested
Theories are generally poorly documented	There are widely used standards for describing theories in a clear and precise way
Difficult to identify the theories that actually are used or have been proposed	For each SE sub-discipline, there are web-sites and systematic reviews that systematize and characterise relevant theories

[D. I. K. Sjøberg, T. Dybå, B. C. D. Anda, and J. E. Hannay. Building Theories in Software Engineering, In: Advanced Topics in Empirical Software Engineering, ed. by Forrest Shull, Janice Singer, Dag I.K. Sjøberg. Springer-Verlag London. (ISBN: 13:978-1-84800-043-8), 2008.]

The need for replication (by others)

Of the 113 controlled experiments, we found 5 close and 15 differentiated replications (other categories of subjects, tasks or systems).

Differentiated replications:

	Same authors	Other authors	Total
Confirmation	7	1	8
Different results	1	6	7
Total	8	7	15

Sjøberg, *et al*. A Survey of Controlled Experiments in Software Engineering, TSE 31(9):733-753, 2005.

So, to make progress, we need to build on the work of each other – we need to share data, testbeds and other artifacts!

Challenges

- How to motivate people to (re)using the material of others?
- How to motivate people to package (raw) data and material, and support others so that they can (re)use it in a satisfactory way (for both parties)?

Table 1. Data/artifact license taxonomy			
Attribute	**Property**	**Value**	**Definition**
Lifetime	Permission	Single use	Can use artifact only for one application
		Limited	Can use artifact repeatedly for a set period of time
		Unlimited	Unlimited use of the artifact
Area	Permission	Specific project	Can use artifact only for this project
		Unlimited	Unlimited use of the artifact
Data	Protection	Sanitized	No personal information contained
		Proprietary	Data contains information that uniquely identifies individuals of specific organizations
Transfer to 3rd party	Permission	No	Only licensee can use artifact
		Yes	Licensee can pass on artifact under the same license conditions applicable to this licensee. This may require a non-disclosure agreement with either this licensee or owner of artifact.
		Yes after time period	Licensee can pass on artifact after a period of time (e.g., artifact is restricted for 3 years then available to anyone)

[V. Basili, M. Zelkowitz, D. I. K. Sjøberg, P. Johnson, and T. Cowling. Protocols in the use of Empirical Software Engineering Artifacts, *Empirical Software Engineering* 12(1):107–119, 2007]

How to improve the quality of SE research?

- Increase competence on how to conduct empirical studies
  - Guidelines and empirical methods included in SE curricula
  - Develop infrastructures to support the conducting of studies
- Improve the links between academia and industry
  - Get involved in SPI work in companies
  - Give seminars and courses where studies are included
- Develop common research agendas
  - More concentrated effort – SE researchers should work on common research programs
- Consult related disciplines
  - SE is typically performed by humans in organisations. Hence, Simula has established collaborations with disciplines such as psychology, sociology and management, in addition to statistics
- Increase the resources available

Large-scale empirical work requires a great amount of resources

- At Simula we used to spend about 25% of budget on empirical studies, mainly at the expense of more researchers.
- In research grants applications, one budgets for money for positions, equipment and travel; why not include money for conducting empirical studies?
- Given the importance of software systems in society, there is no reason why research projects in SE should be less comprehensive and cost less than large projects in other disciplines, such as physics and medicine. The U.S. funding for the Human Genome Project was $437 million over 16 years. If related activities are included, the total cost rises to $3 billion! CERN's annual budget is about $800 million.
- An ambitious, long-term goal would be to establish a research programme in SE similar to the Human Genome Project.

Summary: “The Simula opportunity”

- Development of a research group almost from scratch
- Strong research management, focused research
- Strong links to industry, both for research and for technology transfer (experiments, case studies, SPI, seminars, courses, etc.)
- The use of resources mostly up to the department:
  - Before, 2/3 of the department’s budget was bound into salaries (more now)
  - Extensive use of professionals to take part in studies
  - Employment tailored to the needs of the group (e.g., professional project and knowledge manager)
  - Use of consultants for research support
  - Development of sophisticated experiment support environments