A Survey on Controlled Experiments in Software Engineering

Lutz Prechelt
prechelt@inf.fu-berlin.de
Institut für Informatik, Freie Universität Berlin
2006-01-05

Source

- Dag I.K. Sjøberg, Jo E. Hannay, Ove Hansen, Vigdis By Kampenes, Amela Karahasanovic, Nils-Kristian Liborg, Anette C. Rekdal: *"A Survey of Controlled Experiments in Software Engineering",*
  IEEE Transactions on Software Engineering 31(9), September 2005.

Scope

- Controlled experiments published in nine journals and three conferences during 1993 to 2002
  - but excluding pure HCI or IS studies
- Journals;
  - ACM Transactions on Software Engineering Methodology (TOSEM)
  - Empirical Software Engineering (EMSE)
  - IEEE Computer
  - IEEE Software
  - IEEE Transactions on Software Engineering (TSE)
  - Information and Software Technology (IST)
  - Journal of Systems and Software (JSS)
  - Software Maintenance and Evolution (SME)
  - Software: Practice and Experience (SP&E)
- Conferences
  - Intl. Conf. on Software Engineering (ICSE),
  - IEEE Intl. Symp. on Empirical Software Engineering (ISESE)
  - IEEE Intl. Symp. on Software Metrics (METRICS)

Prior Surveys

	Our survey
Purpose	Surveys topics, subjects, tasks, environments, and internal and external validity of controlled experiments in SE
Scope	SE
Journals	EASE, EMSE, ICSE, IEEE Computer, IEEE Software, ISESE, IST, JSME, JSS, METRICS, SP&E, TOSEM, TSE
Sampling of papers	All papers in the period 1993-2002
Number of investigated papers	5453 papers scanned, 103 papers analysed in depth

Prior Surveys (2)

There are a few other surveys

- on narrower topics, e.g.
  - object-oriented technology (Deligiannis, Shepperd, Webster, Roumeliotis 2002)
  - testing techniques (Juristo, Moreno, Vegas 2004)
  - software effort estimation (Jørgensen, Teigen, Moløkken 2004)
- on only one single conference
  - Shaw on ICSE 2002
- or with theory-formulating intent
  - Zendler on 31 controlled experiments (2001)

Definition: Controlled Experiment

- Experiment:
  - A study in which an intervention is deliberately introduced to observe its effects
- Controlled Experiment:
  - Comparing the effects of two (or more) different interventions, while keeping all other conditions constant
  - Note: If people are involved, constancy usually requires
    - multiple trials for each condition and using averaging,
    - random assignment of subjects to interventions
  - However, the survey also includes *quasi-experiments*, in which random assignment is missing

Number of articles per source

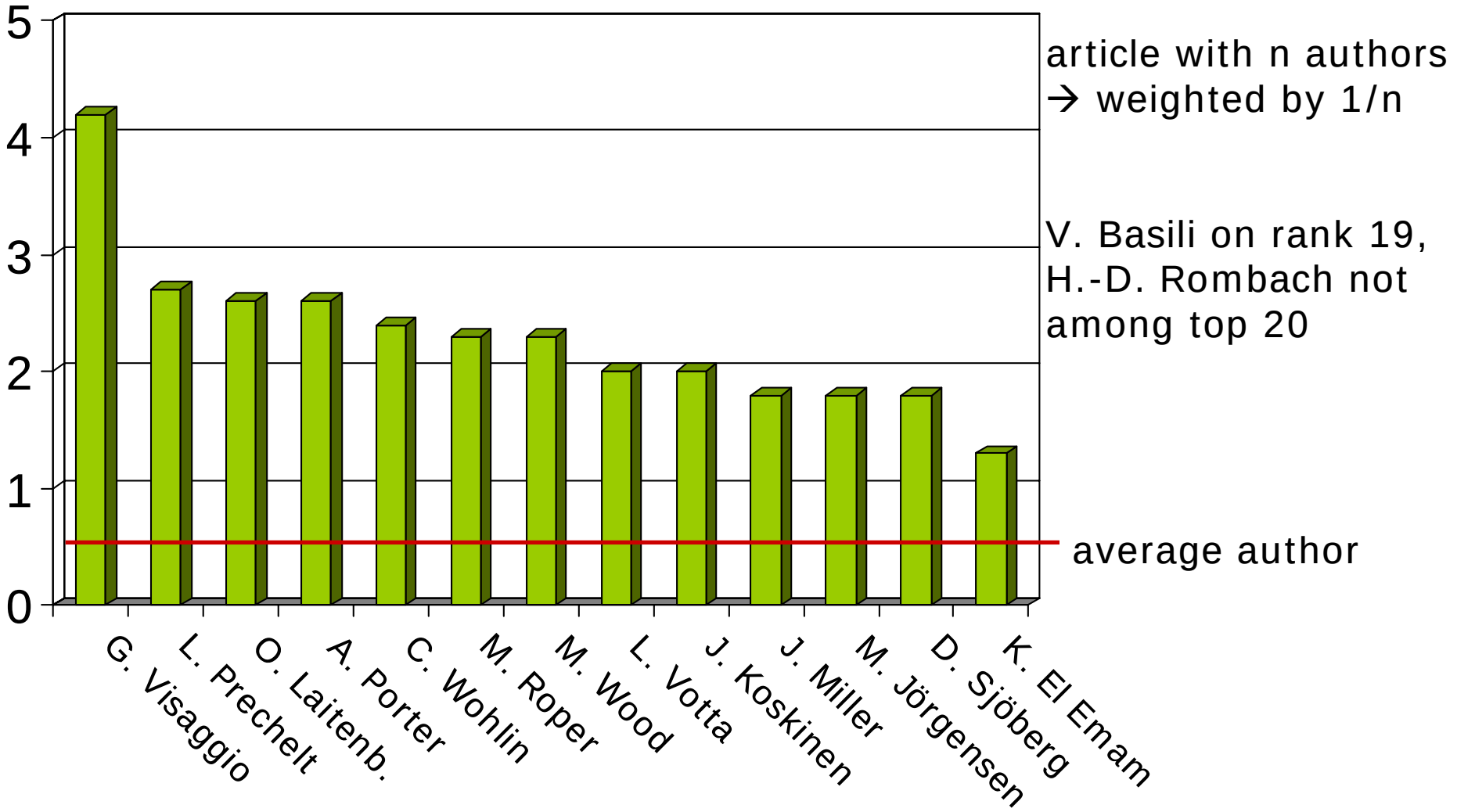
Journal/Conference	articles investigated	reporting controlled experiments	
		N	Row %
EMSE	124	22	17.7
ISESE	20	3	15.0
METRICS	177	10	5.6
JSS	886	24	2.7
TSE	687	17	2.5
ICSE	520	12	2.3
IST	745	8	1.1
SME	186	2	1.1
IEEE SW	532	4	0.8
TOSEM	125	1	0.8
IEEE Comp	780	0	0
SP&E	671	0	0
All	**5453**	**103**	**1.9**

Number of experiments

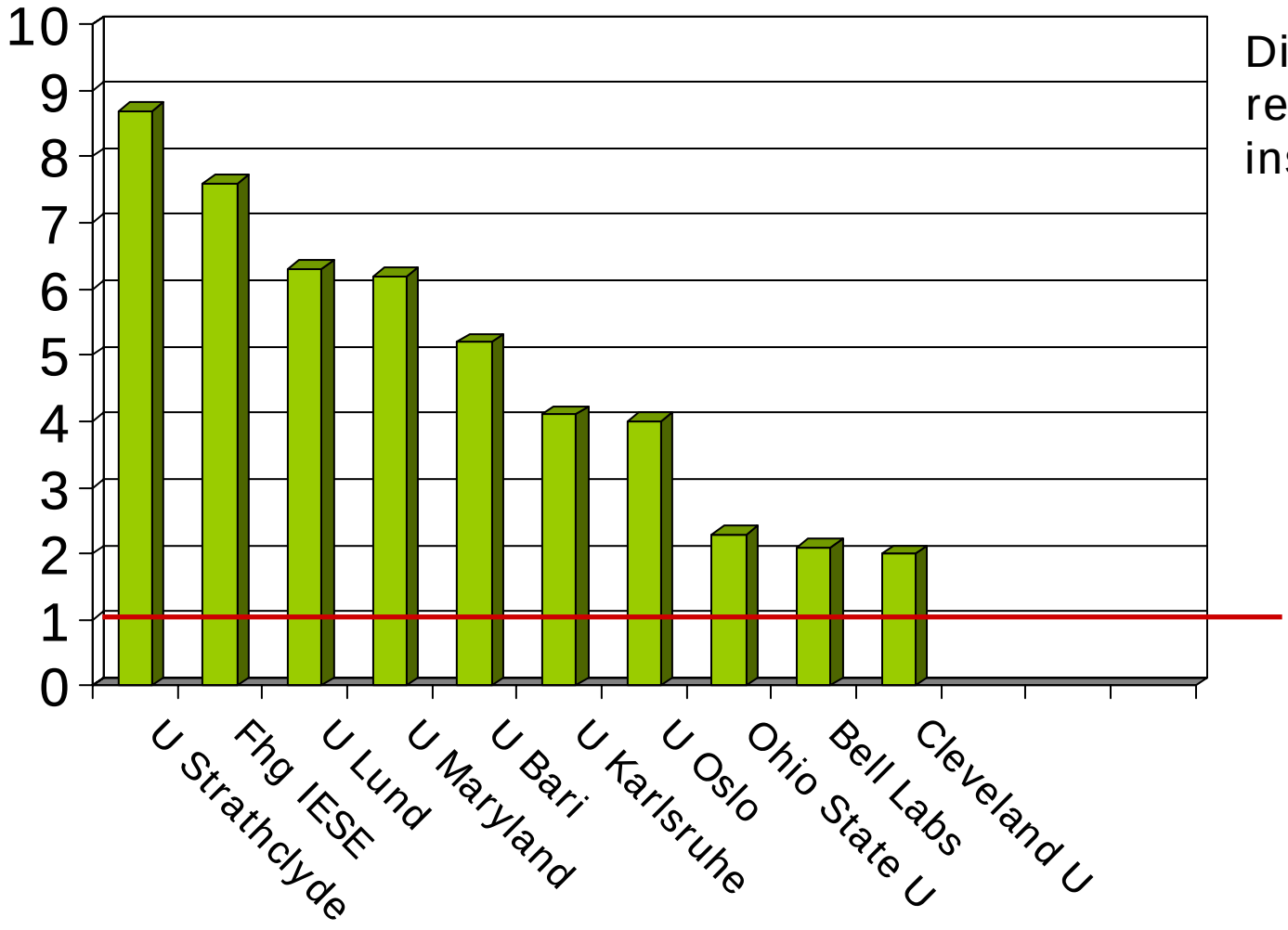
- 103 articles describing experiments
  - 207 scientists from 109 institutions in 19 countries are involved as authors
- 12 of the articles report more than one experiment
- 4 experiments occur in more than one article

- Density of experiments across articles:
  - Tichy et al. survey: 14% of articles in SE journals describe empirical work of some kind
  - Glass et al. survey: 3-4% describe controlled experiments
  - Zelkowitz/Wallace survey: 3% describe controlled exp.
  - Here: 2% describe controlled experiments

Who did them: Individuals



Who did them: Institutions



Distorted, because researchers switch institutions

Topics of articles

- The experiment topics cover a broad range
- But most topic areas are represented by only 1 or 2 articles
  - 34 areas are mentioned overall
- Notable exceptions:
  - Inspections and reviews: 35 articles
  - OO design methods: 8 articles
  - Process models: 5 articles

Subjects

Subject Category	N	%
Undergraduates	2969	54.1
Graduates	594	10.8
Students, type unknown	1203	21.9
Professionals	517	9.4
Scientists	74	1.3
Unknown	131	2.3
Total	**5488**	**100**

- 7 articles are on experiments where both students and professionals were present
- Only 3 of them assess the differences. Findings:
  1. no difference
  2. professionals better
  3. 3 tasks: no difference in two, professionals better in the third

Subject mortality

- Mortality is the dropping-out of subjects during the experiment
  - No or no complete results can be reported for these subjects
  - If mortality depends on the intervention (group membership), the effect damages the random assignment and threatens the experiment's validity
- Mortality is reported for only 24 experiments
  - and for them was 2% on average

Information about the subjects

- "In order to generalize from an experiment [...] one needs information about various characteristics and their variation both in the sample and in the group to which the results will be generalized (target population).
- [...] However, there is no generally accepted set of background variables for guiding data collection in a given type of study, simply because the software engineering community does not know which variables are the important ones."

Information about the subjects (2)

Freie Universität Berlin

- 91 articles on experiments with students reported
  - gender: for 7 experiments
  - age: 6
  - grades: 6
  - programming experience (general): 17
  - prog. exp. (years, #langs): 11
  - has industrial work experience: 9
  - years of industrial work exp.: 9
  - task-related experience: 64
  - task-related training: 27
- 27 articles on experiments with professionals reported:
  - gender: 2
  - age: 3
  - job type: 7
  - degrees: 3
  - programming experience (general): 2
  - prog. exp. (years, #langs): 7
  - prog. exp. self-assess: 2
  - years of industrial work exp.: 5
  - task-related experience: 14
  - task-related training: 12

Information about the subjects (3)

Examples of reports on programming experience:

- Negative example:
  - "Some of the students had industrial programming experience."
- Positive example:
  - "On average, subjects' previous programming experience was 7.5 years, using 4.6 different programming languages with a largest program of 3510 LOC. Before the course, 69 percent of the subjects had some previous experience with object-oriented programming, 58 percent with programming GUIs"

Subject recruitment and compensation

Freie Universität Berlin

- Subjects participation was
  - mandatory: 12 student experiments
  - voluntary: 29 experiments
  - not reported: 72 experiments

- Student subject compensation was
  - better grades: 10 experiments
  - extra credits: 9 experiments
  - money: 3 experiments
  - exhibition trip: 1 experiments
  - no reward: 16 experiments
  - nothing reported: 74 experiments

Subject sampling

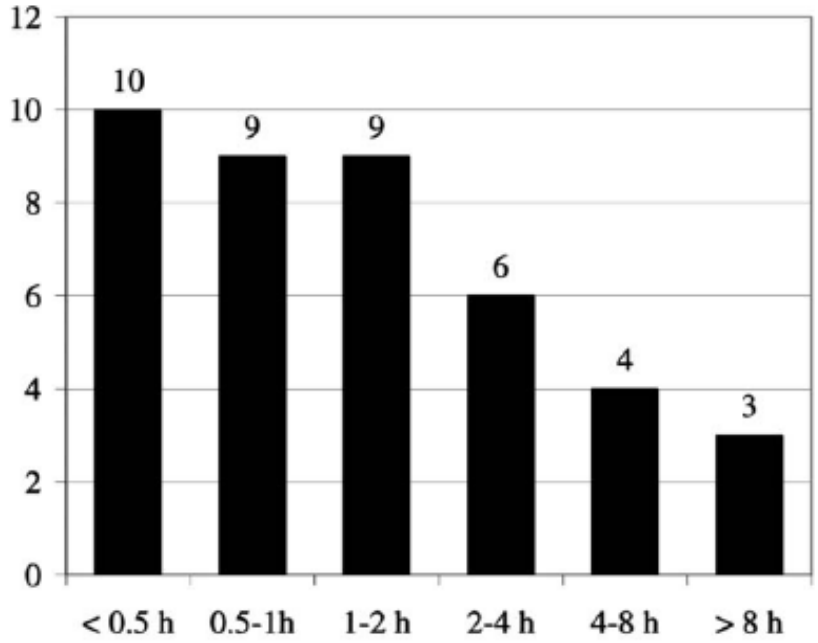
- Only 1 article explicitly reported random sampling from a defined target population
- A few claimed random sampling
  - but did not report population or procedure
- By far most studies used convenience sampling
  - which may introduce bias

Task type and duration

Tasks are classified into four broad categories:

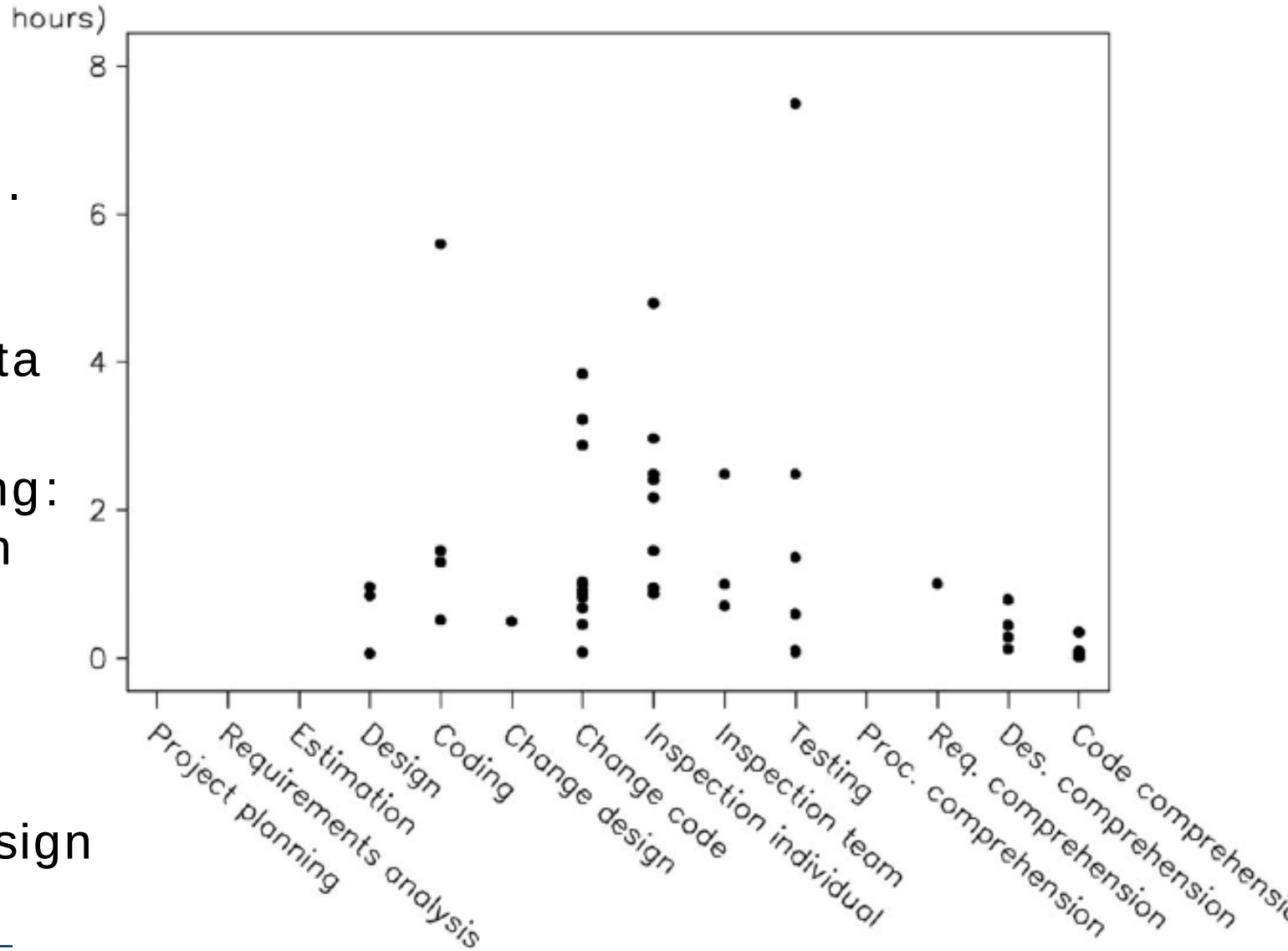
- Plan (plan, estimate, negotiate) 10%
- Create (design, code) 20%
- Modify a design or code 16%
- Analyze (inspect (37%), test/debug, comprehend) 54%

- Work time per subject is reported for only 41 exp.
  - See figure for distribution of median work times:
  - Two thirds of all tasks have a median duration of two hours or less



Task duration by type

- the 41 experim. with detailed time data only
- 4 missing: 18, 55 h Coding, 25 h change code, 18 h design



Programs/materials under study

- Only 16 experiments expressedly use 'real' applications
  - 10 for inspection tasks
  - 2 for design tasks

Application type	N	%
Constructed	80	70.8
Commercial	16	14.2
Student project	5	4.4
Open source	0	0.0
Unclear	12	10.6
Total	**113**	**100**

- Only 67 experiments report on size of materials
  - range from 1 page to 67 pages for documents
  - range from 25 LOC to 3955 LOC for code
  - partially dependent on task type, e.g. larger for inspections

Experiment location

- classroom setting 10
- laboratory 20
- "academic" setting 9
- real office 1
- not reported 73

- Only about half report the name of the institution (university, company)

Experiment equipment

Tool	N	%
PC or workstation (only)	32	28.3
Pen and paper (only)	25	22.1
Combination	5	4.4
Unknown	51	45.1
Total	**113**	**100**

- Almost half of the experiments do not report on the tools used by the participants!
  - Most of these are probably pen-and-paper
- Pen-and-paper may be OK for inspections, but is dubious for most other types of task

Replication

- 20 of the exp. call themselves replications of others
  - 7 of them in inspections, 5 in maintenance
- 5 are close replications, 15 are differentiated
  - 4 with different programs, 3 with different tasks, many with different kinds of subjects
- All 5 close replications confirm the original results
- Of 7 differentiated repls. performed by other authors , only 1 confirms the original results
- Of 8 differentiated repls. performed by other authors , 7 confirm the original results

Threats to internal validity

Category	Threat not handled	Threat reduced	Threat eliminated	Total
Selection	10	35	7	**52**
Instrumentation	9	30	6	**45**
Maturation	3	14	6	**23**
Testing	2	22	4	**28**
History	3	9	6	**18**
Attrition	5	3	4	**12**
Regression	0	1	1	**2**
Ambiguous Temporal Precedenc	0	0	0	**0**
Additive and Interactive Effects	0	0	0	**0**
No of threats*	**32 (17.8%)**	**114 (63.3%)**	**34 (18.9%)**	**180 (100%)**
No of Experiments	**26 (23.0%)**	**55 (48.7%)**	**19 (16.8%)**	**71† (62.8%)**

- Many articles fail to discuss internal validity
- About half of all experiments appear to be quasi-experiments only (selection effects)

Threats to external validity

Factors addressed as threats to external validity	Experiments	%
Subject (only)	14	12.4
Task (only)	10	8.8
Environment (only)	1	0.9
Subject and environment	2	1.8
Subject and task	31	27.4
Subject, environment and task	14	12.4
Treatment and subject, task or environment	6	5.3
Threats to external validity not addressed	35	31.0
Total	**113**	**100**

- More than half do not contain adequate discussion
  - none at all or one area of concern only
- Few consider the environment or the treatment
- Reporting is rather unsystematic

Conclusions

- The analyses performed in such a survey cannot guarantee that an article is good
  - "Good" meaning credible and reliable
- However, many of its aspects can point out where an article is not good
- The average quality in these respects is not very good
  - so it is not really difficult to be in the upper half
  - but there are quite a number of issues to think of!

- Want to learn how to do good experiments?
- Come to "Empirische Bewertung in der Informatik" (2V+2Ü, 6 LP, SoSe)

Thank you!