28th International Conference
on Software Engineering

On the Success of Empirical Studies in ICSE

(What should be changed?)

Carmen Zannier, Grigori Melnik, Frank Maurer
University of Calgary, Canada

ebe.cpsc.ucalgary.ca/ebe

The State of the Art …

Motivation & Context

- Research question:

  Are we getting better in conducting empirical research in SE?

- Context: ICSE

Hypotheses

H_1: The **quantity** of empirical evaluations performed has increased over 29 years of ICSE proceedings

H_2: The **soundness** of empirical evaluations has improved over 29 years of ICSE proceedings

- very basic criteria for soundness:
  - where appropriate, well-defined hypotheses stated
  - 4 parameters:
    - *Study Type*
    - *Sampling Type*
    - *Target and Used Populations (do these match?)*
    - *Evaluation Type (self-confirmatory / independent)*
  - legal (proper) use of a method of analysis

Sampling

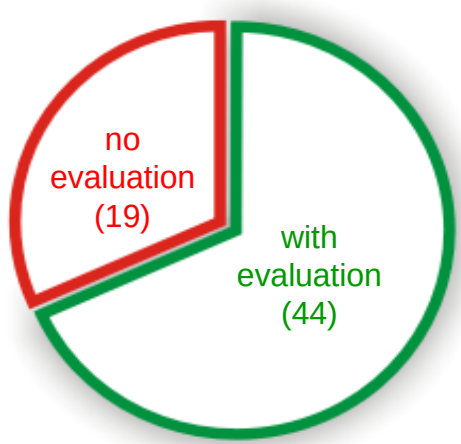
- Target population – all accepted peer-reviewed ICSE publications (technical papers and experience reports) (N = 1227)
- Used population – accepted peer-reviewed ICSE publications
- Sample – stratified random (n = 63) into 9 clusters of 3 papers (covering 3 year periods)
- For Hypothesis 1, we used 2 groups:
  - early years of ICSE (1974-1990)
  - later years of ICSE (1991-2005)

Procedure

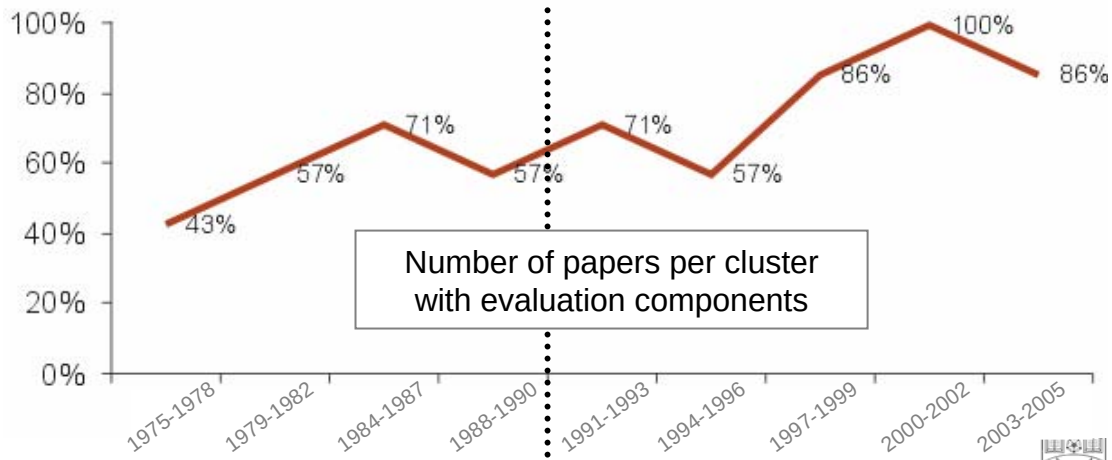
- Independent evaluation
- Investigator 1 examined each paper from the sample
  - contain evaluation?
  - is evaluation sound?
- Analysis replicated internally
  - blind evaluation by investigator 2 and 3
    - resulted in validation and more precise formulation of study types
    - 12/63 randomly assigned papers used for validation

Results – Hypothesis 1 (Quantity)

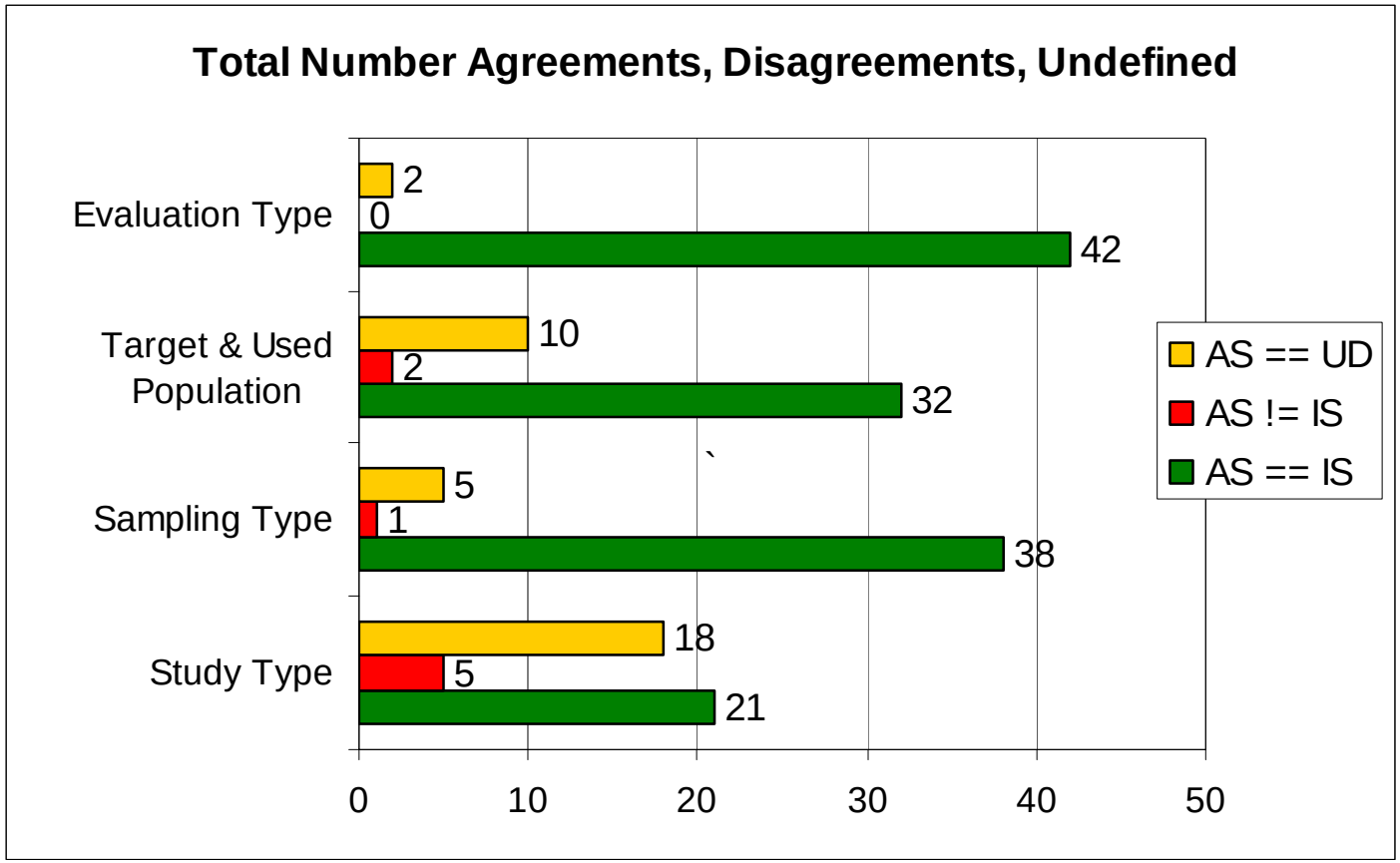
- Null hypothesis is rejected at level 0.05
- Thus, we conclude that **empirical evaluation in software engineering field is becoming more common**.
- Maturation of the field?



ICSE lifespan (27 conferences)



Results – Clarity of Evaluation Properties

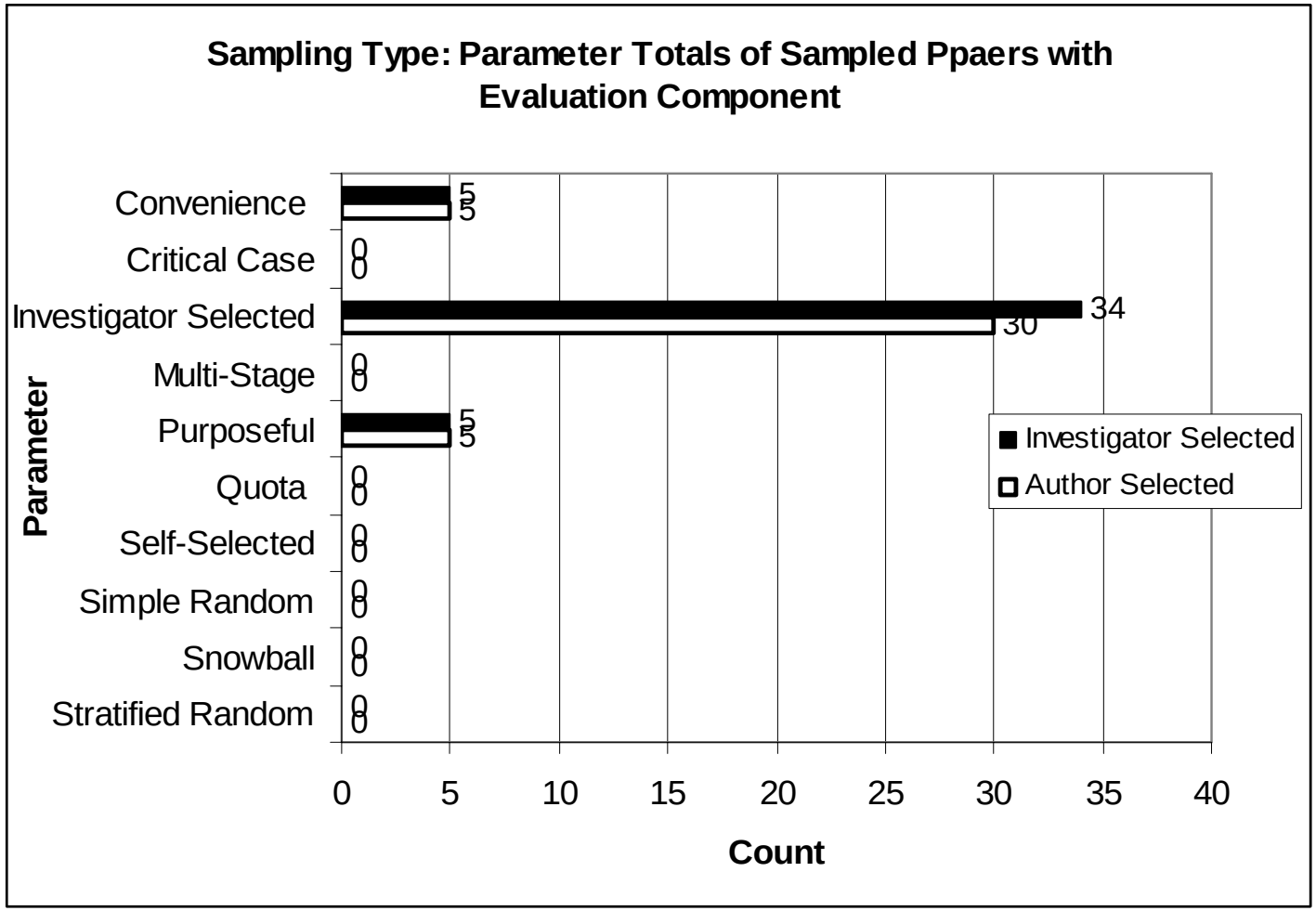


AS = Author-selected, IS = Investigator-selected, UD = undefined

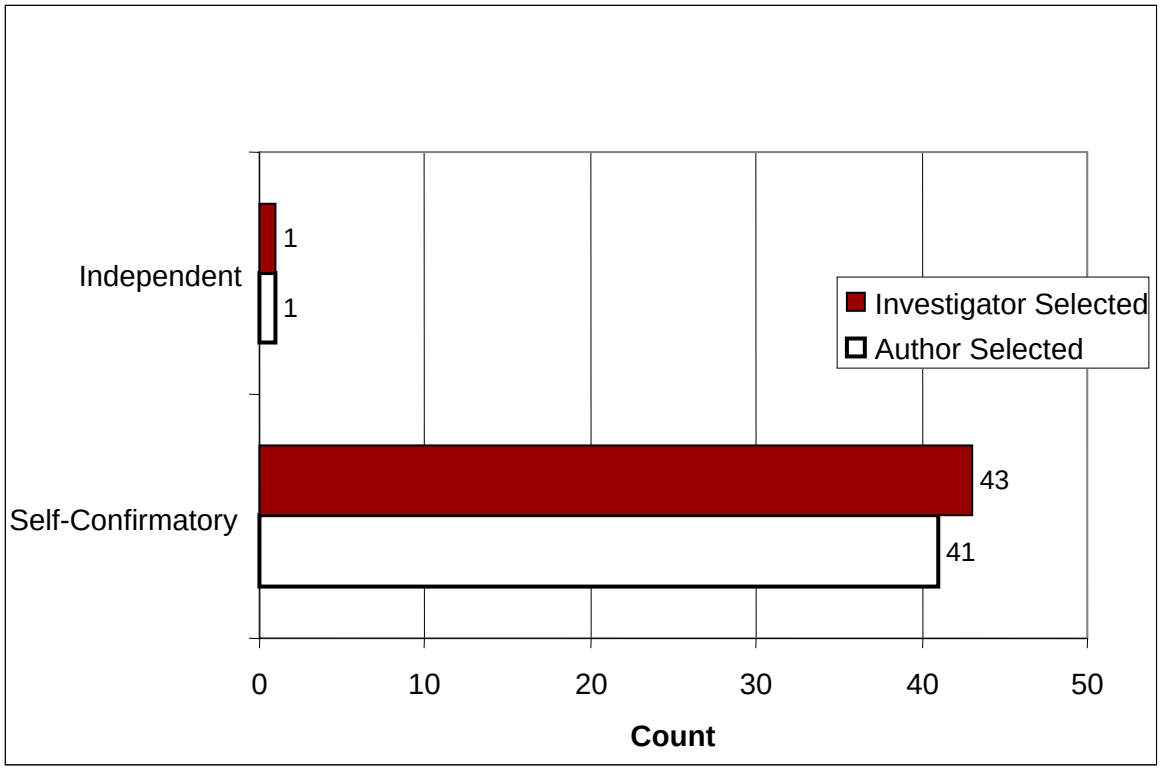
Hypotheses are not stated explicitly

- Except for 1 study in our random sample, none of the examined studies that should have stated hypotheses or propositions (17) contained hypotheses clearly stated
- This is despite published guidelines and numerous recommendations by empirical experts

Results – Sampling Type



Results – Major Concern 1



Though we've seen more empirical studies done, most of them are self-confirmatory in nature!

⇒ 5 criteria of soundness are not improving,

⇒ Hypothesis 2 is rejected (qualitatively)

Results – Major Concern 2

- Replication anyone?

- Possible reasons:
  - ICSE reviewers consider the "excitement factor" to draw the crowd and replicated studies may not rank highly unless they are contradicting some known data
  - Replicated studies are simply not being done
    - similar situations in journals (informal evaluation)

Towards a Family of Studies of Quality of ESE

- A Survey of Controlled Experiments in Software Engineering by Sjøberg et al, J.TSE, 31(9): 733-753, 2005
  - controlled experiments over last 10 years
  - scope: major conferences + journals + magazines
  - findings are similar to ours:
    - prevalence of academic studies,
    - hypothetical (not real-world) applications,
    - reporting "vague and unsystematic"
  - another problem - a lack of consistent terminology

What Does This Mean to Us as a Community ?

- Researchers recognize a need for empirical evaluation to get their papers accepted (at ICSE)
- However, the soundness of empirical evaluations has not improved over 29 years of ICSE
  - self-confirmatory studies dominate
  - no replication
  - hypotheses are not specified explicitly
  - non-random sampling
  - inconsistent definition of study types
- Does this mean we (researchers) are paying a lip service to empirical evaluation?

Food for Thought

- Problem space is so huge, that it is hard to come up with valid results in industrial time frames feasibly
- "Over the years I've matured from quantitative to qualitative" (Vic Basili)