Are we Getting Reliable Evidence?

Methodology is Critical in Empirical Studies

Natalia Juristo
Universidad Politécnica de Madrid

Project PID2022-137846NB-B-I00 funded by:

Schedule

Empirical Research

- What is empiricism??
- Types of empirical studies

Reliability and Methodology

- What is reliability?
- What is methodology?

Cases

- Status of methodology quality in some SE cases

Conclusions

Schedule

Empirical Research

- What is empiricism??
- Types of empirical studies

Reliability and Methodology

- What is reliability?
- What is methodology?

Cases

- Status of methodology quality in some SE cases

Conclusions

Empirical Research

Empiricism **collects data** from reality to gain knowledge

Rationalism is good but not enough for gaining knowledge

Types of Empirical Studies

OBSERVATIONAL

INTERVENTIONAL

Observational study

No intervention

In a natural uncontrolled environment

Field experiments

Purposely intervention

Natural uncontrolled environment

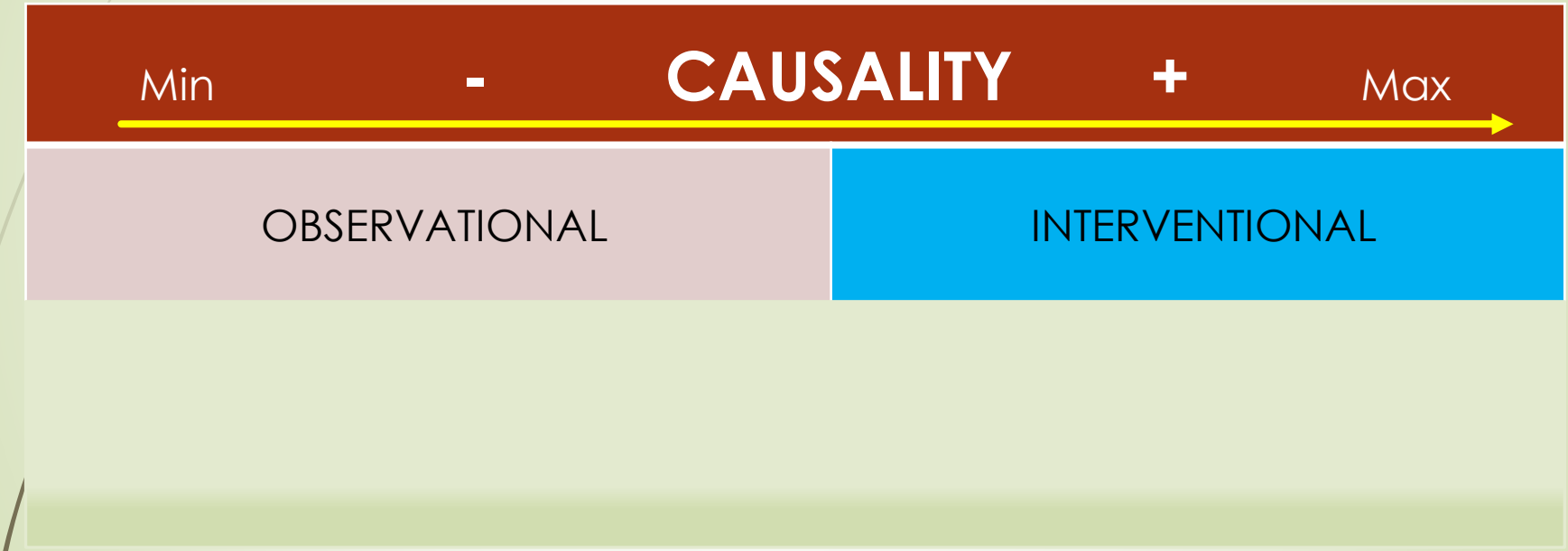
Clinical Trials

Purposely randomized intervention

Natural uncontrolled environment


Laboratory experiments
Purposely intervention
Highly controlled artificial environment

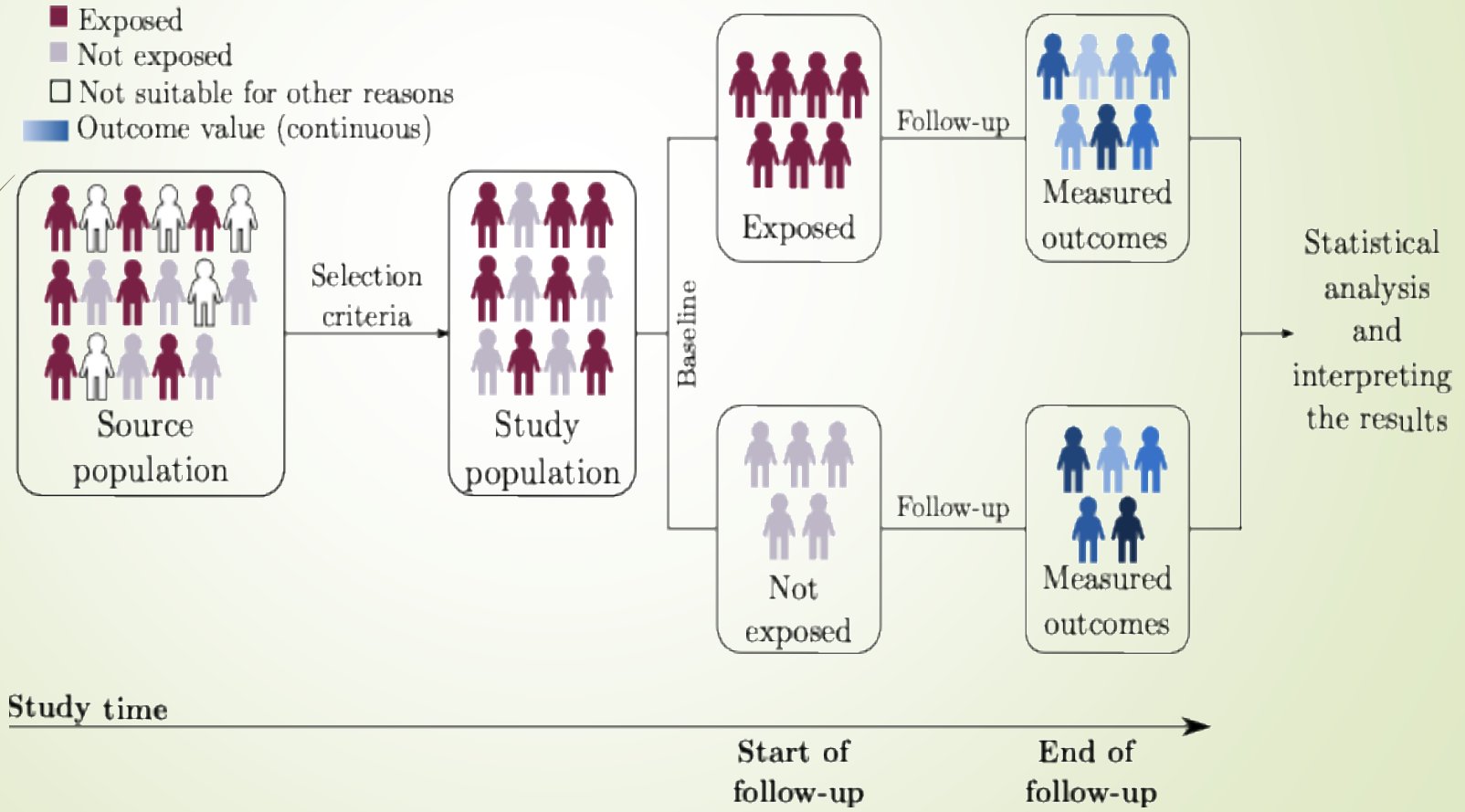
Types of Empirical Studies




Observational study
collecting data from reality

Retrospective Observational Study

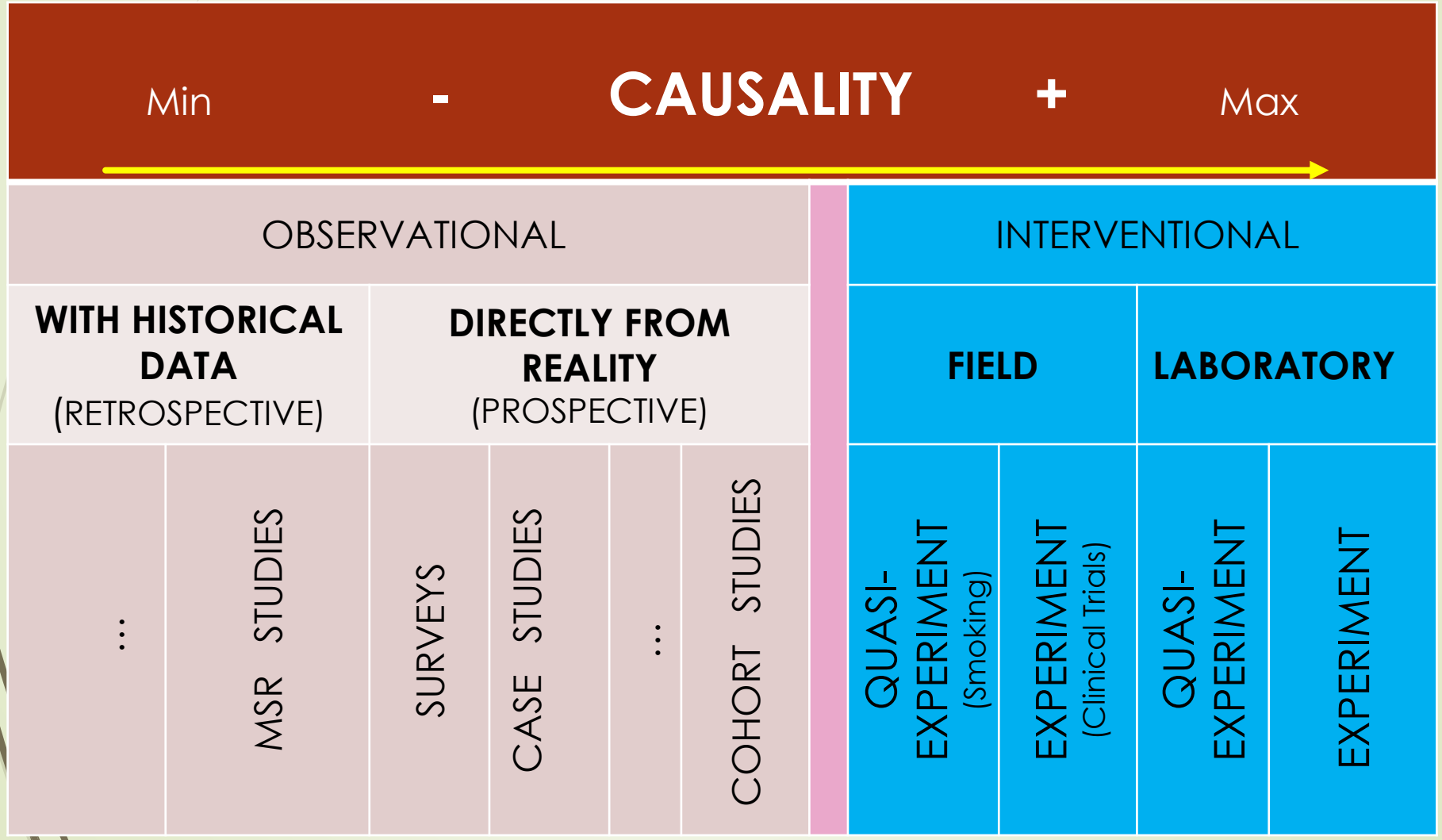
working with data rather than with reality



Causality: Intervention + Control

- Intervention
  - **Decouples** the variable under study from other unknown linked variables
  - Get **a chance** to reveal causality between the manipulated variable and the measure effect
- Control
  - Uninteresting but known variables can be controlled
    - Fix or **Blocking**
  - Strategies can be applied to avoid the influence of uninteresting variables on studied output
    - Unknown variables: **Randomization**
  - Etc.

Types of Empirical Studies



Schedule

Empirical Research

- What is empiricism??
- Types of empirical studies

Reliability and Methodology

- What is reliability?
- What is methodology?

Cases

- Status of methodology quality in some SE cases

Conclusions

Types of Results

- The goal of conducting an empirical study is obtaining evidence

- Each type of empirical study obtains different type of evidence
  - **Lab experiments** reveal causal relationships but with low external validity
  - **Field experiments** are weaker in detecting causality but stronger in external validity
  - **Observational prospective** studies might arrive closer to causality if extraneous variables are controlled
  - **Observational retrospective** studies reveal correlations
  - **Case studies** reveal anecdotes
  - **Surveys** reveal opinions

Results Reliability

- Extent to which an empirical result **corresponds** accurately to the **real world**

- Just running an empirical study does not guarantee reliable results

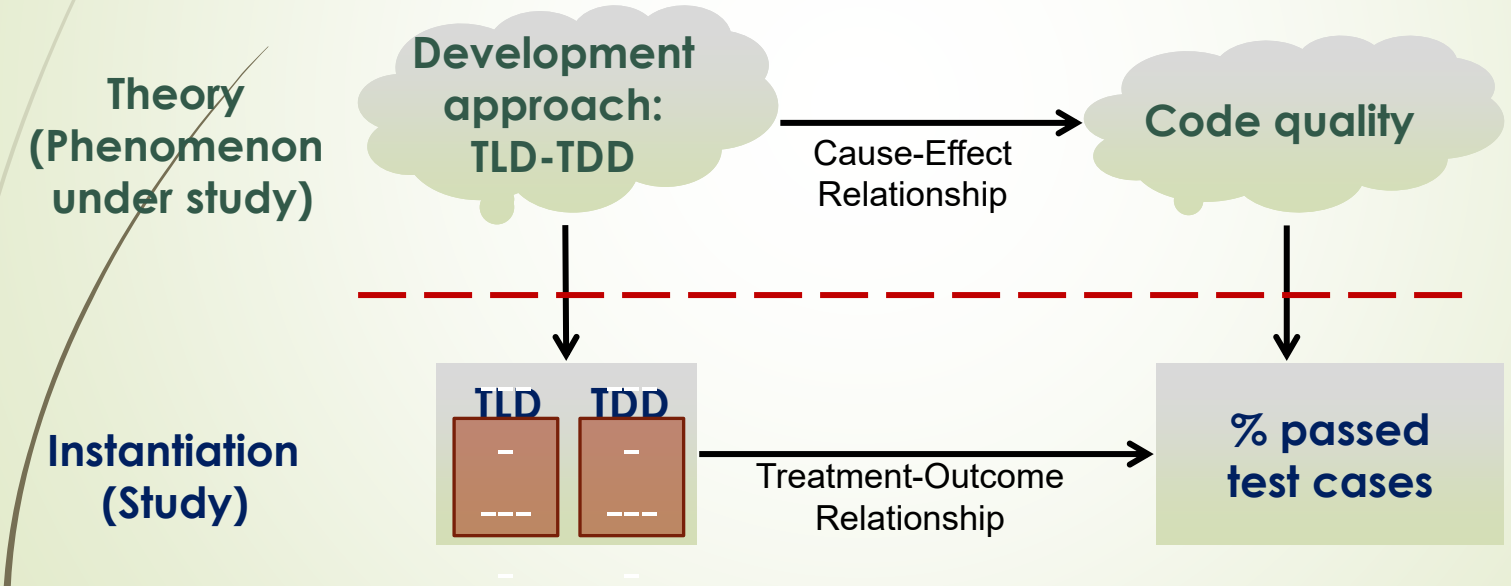
Methodology

- A systematic approach that describes all steps throughout the process of conducting an empirical study
- Methodology rules how to proceed for each step
  - plan, design, collect data and analyze data
- Different types of empirical studies keeps up with different methodology/procedure
  - What needs to be done in each step for experiments, observational studies, case studies, etc. is different

Reliability & Methodology

- **Proper adherence** to methodology supports results reliability
  - On trustworthy findings, evidence can be built
  - Untrustworthy findings are useless or even misleading
- **Methodological rigor** in carrying out empirical studies is essential to obtain trustable findings
  - Methodology helps increasing study validity
    - Conclusion and Internal validity

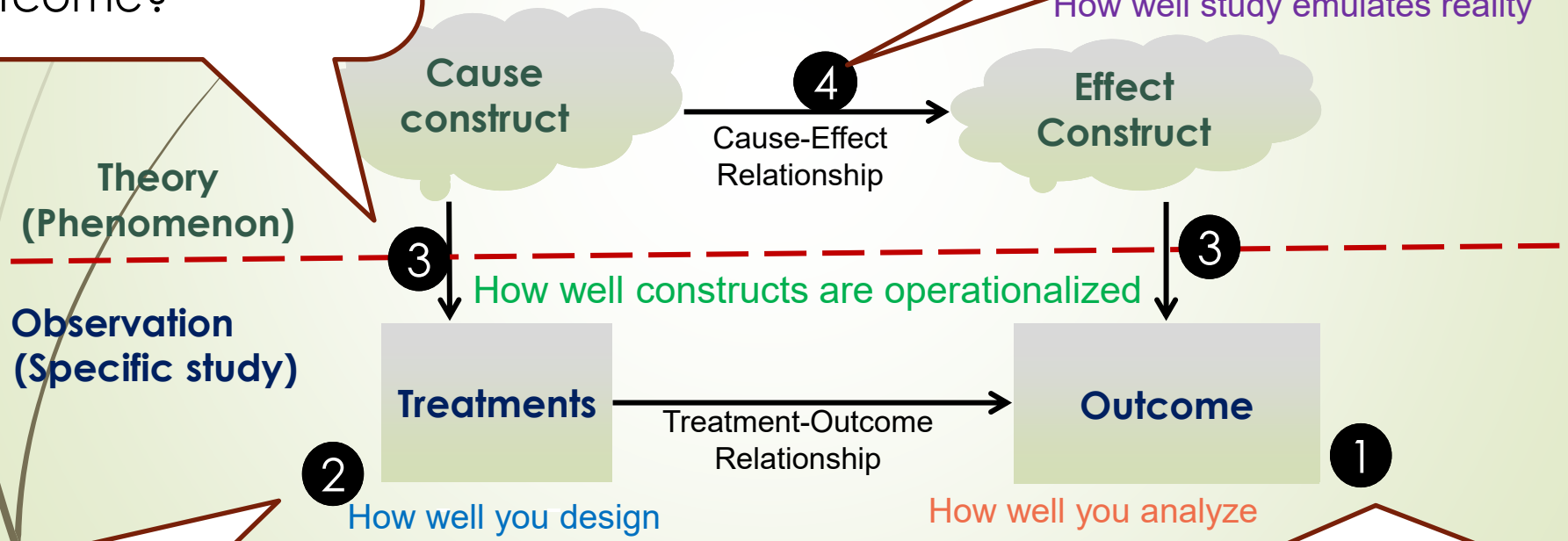
On Study Validity



Types of Validity

Construct validity
Proper operationalization of treatments and outcome?

External validity
Application of study results to reality?



Internal validity Proper control of extraneous variables?

Conclusion validity Proper application of data analysis techniques?

Schedule

Empirical Research

- What is empiricism??
- Types of empirical studies

Reliability and Methodology

- What is reliability?
- What is methodology?

Cases

- Status of methodology quality in some SE cases

Conclusions

Some cases

- Experiments
  - Analysis of experiments (2006 - 2015)
  - Analysis of cross-over designs (2012 - 2014)
  - Aggregating families of experiments (1998 - 2016)
- Software repositories
  - Use of term experiment (2015 - 2018)
  - Awareness of study type (2020 - 2022)

Are experiments properly Analyzed?

2018 ACM/IEEE 40th International Conference on Software Engineering

Statistical Errors in Software Engineering Experiments: A Preliminary Literature Review

Rolando P. Reyes Ch.
Universidad Politécnica de Madrid
Madrid, Spain
Universidad de las Fuerzas Armadas ESPE
Sangolquí, Ecuador
rpreyes1@espe.edu.ec

Oscar Dieste
Universidad Politécnica de Madrid
Madrid, Spain
odieste@fi.upm.es

Efraín R. Fonseca C.
Universidad de las Fuerzas Armadas ESPE
Sangolquí, Ecuador
erfonseca@espe.edu.ec

Natalia Juristo
Universidad Politécnica de Madrid
Madrid, Spain
natalia@fi.upm.es

ABSTRACT

Background: Statistical concepts and techniques are often applied incorrectly, even in mature disciplines such as medicine or psychology. Surprisingly, there are very few works that study statistical problems in software engineering (SE). ***Aim:*** Assess the existence of statistical errors in SE experiments. ***Method:*** Compile the most common statistical errors in experimental disciplines. Survey experiments published in ICSE to assess whether errors occur in high quality SE publications. ***Results:*** The same errors as identified in others disciplines were found in ICSE experiments, where 30% of the reviewed papers included several error types such as: a) missing statistical hypotheses, b) missing sample size calculation, c) failure to assess statistical test assumptions, and d) uncorrected multiple testing. This rather large error rate is greater for research papers

KEYWORDS

Literature review, Survey, Prevalence, Statistical errors

1 INTRODUCTION

Experimentation makes extensive use of statistics. Several studies warn about the existence of scientific articles using inappropriate statistical procedures [5, 32, 62]. This happens even in mature disciplines, such as the health sciences [6].

In turn, there are very few papers studying statistical errors in software engineering (SE) articles. There are papers discussing statistical power [23], heterogeneity in meta-analysis [52], and the relative strengths and weaknesses of cross-over designs [41, 81]

Study

- Dates
  - 2006 – 2015
- Venues
  - Conferences: ICSE

Analysed Papers

- 21 stand alone experiments
- 30 experiments as evaluations

Year	Total papers (TP)	After screening	Selected	%
2006	72	8	2 (3)	2.8% (4.1%)
2007	64	7	2 (3)	3.1% (4.7%)
2008	85	8	1 (3)	1.2% (3.5%)
2009	70	7	0 (3)	0.0% (4.3%)
2010	62	5	1 (3)	1.6% (4.9%)
2011	62	5	1 (3)	1.6% (4.9%)
2012	87	31	4 (3)	4.6% (3.5%)
2013	85	8	1 (3)	1.2% (3.5%)
2014	99	11	5 (3)	5.0% (3.1%)
2015	84	11	4 (3)	4.8% (3.5%)
Total	770	101	21 (30)	2.7% (3.9%)

Results

QUESTIONS	Standalone			Section		
	Yes	No	N/A	Yes	No	N/A
Are descriptive statistics, such as means and counts, reported?	95%	0%	5%	77%	13%	10%
Have the test assumptions (i.e., normality and heteroskedasticity) been checked or, at least, discussed	62%	33%	5%	13%	20%	67%
Do researchers calculate and discuss power?	86%	9%	5%	37%	0%	63%
Is multiple testing, e.g., Bonferroni correction, reported and accounted for?	10%	71%	19%	3%	27%	70%

Summary of Results

- At ICSE, in papers which goal is reporting an experiment
  - **Basic** things (as descriptive statistics)
    - **properly** done
  - **Medium issues** (as tests assumptions) critical for conclusion validity but **easy to deal with**
    - **30%** of papers fail
  - **Advanced** issues (as Bonferroni correction) critical for conclusion validity
    - **70%** papers fail
- Notice the involvement of reviewers!!!
  - So a **community** shortcoming ☹

Are Crossover Experiments Properly Analyzed?

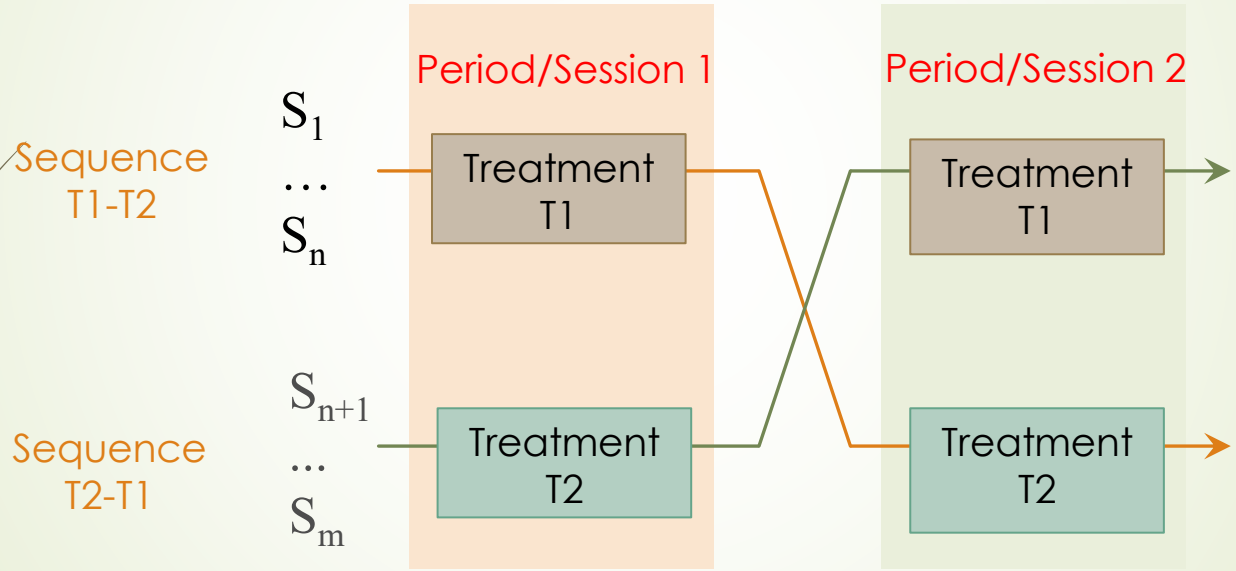
Crossover Designs in Software Engineering Experiments: Benefits and Perils

Sira Vegas, Cecilia Apa, and Natalia Juristo

Abstract—In experiments with crossover design subjects apply more than one treatment. Crossover designs are widespread in software engineering experimentation: they require fewer subjects and control the variability among subjects. However, some researchers disapprove of crossover designs. The main criticisms are: the carryover threat and its troublesome analysis. Carryover is the persistence of the effect of one treatment when another treatment is applied later. It may invalidate the results of an experiment. Additionally, crossover designs are often not properly designed and/or analysed, limiting the validity of the results. In this paper, we aim to make SE researchers aware of the perils of crossover experiments and provide risk avoidance good practices. We study how another discipline (medicine) runs crossover experiments. We review the SE literature and discuss which good practices tend not to be adhered to, giving advice on how they should be applied in SE experiments. We illustrate the concepts discussed analysing a crossover experiment that we have run. We conclude that crossover experiments can yield valid results, provided they are properly designed and analysed, and that, if correctly addressed, carryover is no worse than other validity threats.

Index Terms—Experimental software engineering, controlled experiment, data analysis, crossover design, carryover

Crossover Design

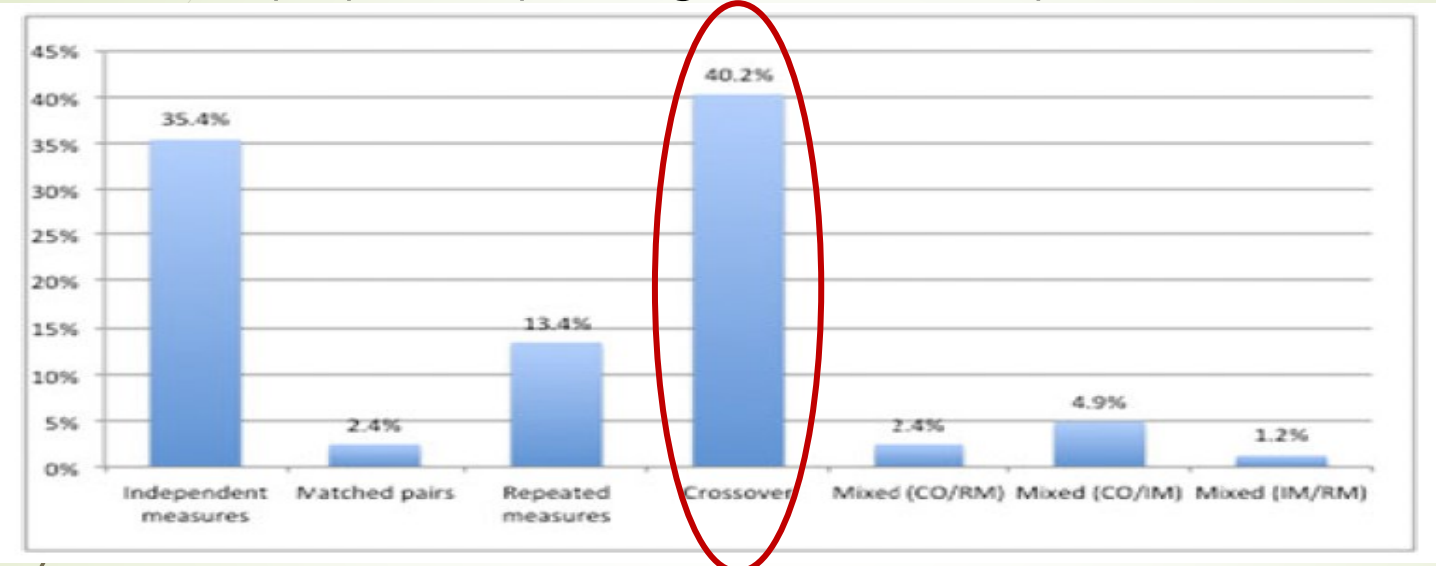


Study

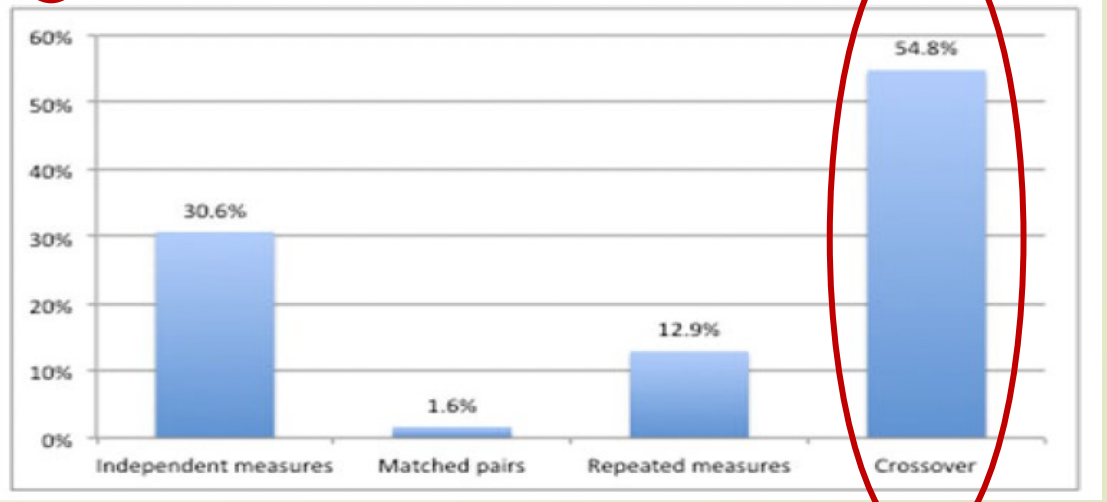
- Dates
  - 2012 – 2014
- Venues
  - Journals: TSE, EMSE, TOSEM
  - Conferences: ICSE, ESEC/FSE, ESEM

Analysed Papers

- 39 papers reporting crossover experiments



- 68 experiments



Results

TABLE 5
Adherence to Good Practices

Good practice	Number of papers adhering to good practice (out of 38)	
Define periods[4]	1	Design
Define sequences	2	
Deal with carryover at design time	0	
Take into account subject variablity	21	Analysis
Deal with carryover at analysis time	0	
Match analysis with design	0	
Beware of effect size[5]	0	

[4]*In this case, there are some papers that address it partially (see details in Section 4).*
[5]*This number is out of 13 papers that calculate effect size. The other 25 papers do not report effect size.*

Summary of Results

- **8%** papers **satisfies some** design practices
- **0%** papers **properly analyse** their crossover design
  - **30%** make a very basic mistake as **choosing a fully unproper technique**

Are Families of Experiments Properly Aggregated?

Analyzing Families of Experiments in SE: A Systematic Mapping Study

Adrian Santos, Omar Gómez, and Natalia Juristo

Abstract—*Context:* Families of experiments (i.e., groups of experiments with the same goal) are on the rise in Software Engineering (SE). Selecting unsuitable aggregation techniques to analyze families may undermine their potential to provide in-depth insights from experiments' results. *Objectives:* Identifying the techniques used to aggregate experiments' results within families in SE. Raising awareness of the importance of applying suitable aggregation techniques to reach reliable conclusions within families. *Method:* We conduct a systematic mapping study (SMS) to identify the aggregation techniques used to analyze families of experiments in SE. We outline the advantages and disadvantages of each aggregation technique according to mature experimental disciplines such as medicine and pharmacology. We provide preliminary recommendations to analyze and report families of experiments in view of families' common limitations with regard to joint data analysis. *Results:* Several aggregation techniques have been used to analyze SE families of experiments, including Narrative synthesis, Aggregated Data (AD), Individual Participant Data (IPD) mega-trial or stratified, and Aggregation of p-values. The rationale used to select aggregation techniques is rarely discussed within families. Families of experiments are commonly analyzed with unsuitable aggregation techniques according to the literature of mature experimental disciplines. *Conclusion:* Data analysis' reporting practices should be improved to increase the reliability and transparency of joint results. AD and IPD stratified appear to be suitable to analyze SE families of experiments.

Index Terms—Family of experiments, meta-analysis, narrative synthesis, IPD, AD

Study

- Dates
  - 1998 – 2016
- Venues
  - Any
  - IST 25%; EMSE 20%; TSE 13%; TOSEM 10%; Others 32%

Analysed Papers

- 39 papers aggregating families of experiments
- 55 aggregations

Results

Technique	N
Narrative synthesis	18
AD	15
IPD mega-trial	13
IPD stratified	6
Aggregation of *p*-values	3

Summary of Results

- **53%** of the families applied **wrong aggregation technique**
- 47% applied a not bad aggregation technique
  - **13%** applied the **proper aggregation technique**

Is the Term Experiment Properly Used at MSR?

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 48, NO. 11, NOVEMBER 2022

Use and Misuse of the Te… in Mining Softwar…

Claudia Ayala

…ware Engineering (SE) has …searchers seem to struggle to …owledge. This is especially the case of MSR …periments and their differences with experiments traditionally …te an experiment from other types of empirical studies and …SR. We analyzed MSR literature obtained from a small-scale systematic …riment in MSR. We found that 19% of the papers claiming to be an experiment are …observational studies, so they use the term in a misleading way. From the remaining 81% of the …to a genuine controlled experiment while the others stand for experiments with limited control. MSR …ook such limitations, compromising the interpretation of the results of their studies. We provide …ons and insights to support the improvement of MSR experiments.

Index Terms—Empirical software engineering, controlled experiment, mining software repositories, research methodology

Now we are entering in the slippery universe of empirical studies with data

Types of Empirical Studies

Min - **CAUSALITY** + Max			
OBSERVATIONAL		INTERVENTIONAL	
RETROSPECTIVE (AKA HISTORICAL)	**PROSPECTIVE** (DIRECTLY FROM REALITY)	**FIELD**	**LABORATORY**

Experiments using repositories

- Data used **as benchmarks to compare** alternatives
  - Alike other **technological experiments** (opposite to human-based experiments)
  - Examples
    - Compare estimation techniques
    - Compare Machine Learning models
    - Compare performance of x algorithms for ranking code search results using data from GitHub projects
- Being experiments they need to follow experiments strategies to control for extraneous variables
  - Randomization, sample selection, etc

Observations using repositories

- Data are used to **extract information** from it
  - Descriptive
    - Descriptive statistics and correlations are extracted from data
    - Only one group -> no comparison
      - Example: Correlation between code smells and defects
    - No control to avoid impact of extraneous variables
  - Analytical
    - Compares 2 (or more) treatments/independent variables to assess the possible effect on an outcome/dependent variable
    - 2 (or more) groups
      - Example: Compares code with code smells vs code without smells regarding the number of defects
    - Need control to avoid impact of extraneous variables with strategies resembling those from experiments

Study

- Dates
  - 2015 – 2018
- Venues
  - EMSE, ESEM and MSR Conference

Analysed Papers

- 35% of the papers use the term experiment
  - 42% at MSR Conference

Venue	Total number of publications per year					Stage 1. Universe of Papers					Stage 2. Primary Studies				
	2015	2016	2017	2018	Total	2015	2016	2017	2018	Total	2015	2016	2017	2018	Total
EMSE	55	68	90	97	*310*	55	68	90	97	*310*	20	36	33	43	*132*
ESEM	63	56	61	58	*238*	63	56	61	58	*238*	8	12	17	11	*48*
MSRConf	72	59	66	77	*274*	42	42	44	48	*176*	18	22	14	20	*74*
Total	190	183	217	232	***822***	160	166	195	203	***724***	46	70	64	74	**254**

Results

MSR Studies using the term experiment	
Observational	47
Interventional	207

MSR Studies using the term experiment		No Hypothesis	No Threats	Use of Causality terms
Observational	47	21	2	3
Interventional	207	74	18	28

Summary of Results

- **18%** papers fully **misusing the term** experiment
  - They are pure observational studies
    - no comparison -> no groups
  - Still **6%** use **causality terms** to discuss results

- **82%** papers **properly use** the term experiment
  - They are interventional
  - With shortcomings
    - For instance, **64%** report **none hypothesis**

What Types of Empirical Study is being Conducted at MSR?

Noname manuscript No.
(will be inserted by the editor)

Empirical Studies in Mining Software Repository. A Review

Nyyti Saarimäki · Fabiano Pecorelli · Valentina Lenarduzzi · Sira Vegas · Gregrey Oko-oboh · Davide Taibi

Received: date / Accepted: date

Abstract *Background. Objective. Method. Results. Conclusion.*

Keywords Cohort Study · Empirical Software Engineering · Empirical Methods

1 Introduction

The common aspect in Mining Software Repositories (MSR) papers is the source of data, not the used methodologies.

Several MSR papers state they do not claim causality. Often causality is not even mentioned.

What are MSR studies doing / not doing preventing them from claiming causality?

Study

- Dates
  - 2020 – 2022
- Venues
  - MSR Conference

Analysed Papers

- 80 papers studied
  - 261 research questions

Year	2020	2021	2022	Total
#papers	44	48	47	139

Year	2020	2021	2022	Total
All papers	44	48	47	139
Excluded	17	18	24	59
Included total	27	30	23	**80**

Results

Study type	# Papers	# RQ
Empirical study	33	112
Not specified	23	51
Experiment	10	32
Case study	7	23
Exploratory study	5	12
Survey	5	8
User study	2	2
Descriptive study	1	6
Evaluation	1	5
Repository mining study	1	2
Interview	1	2
Replication study	1	3
Qualitative study	1	3
Total	91	261

Table 4: Study types claimed by the authors (RQ_1)

Results
Use of Term Experiment

2020 -2022		Number of RQ
Proper Use	Claimed Experiments	35
	Really Experiments	31

Summary of Results

- **25%** specify **none type**
  - 20% of the RQ
- **36%** use the term **empirical study**
  - 43% of the RQ

Schedule

Empirical Research

- What is empiricism??
- Types of empirical studies

Reliability and Methodology

- What is reliability?
- What is methodology?

Cases

- Status of methodology quality in some SE cases

Conclusions

Summaryzing

	Topic	Years	Proper	Innapropriate
Experiments	Checking of test assumptions	2006-2015	70%	**30%**
	Analysis of crossover design	2012-2014	0%	**100%**
	Aggregation of families of experiments	1998-2016	13%	**53%**
Repositories	Use of term experiment	2015-2018	82%	**18%**
	Type of empirical study	2020-2022	Not checked	**61%**

Some Advice

REVIEWERS	• Try to go beyond nicely written paragraphs and truly understand what is being done • Do not object when other reviewers point out serious issues
RESEARCHERS	• Do not copy and paste from previous research • What worked in one study may not work in another
ESE COMMUNITY	• Should work harder on disseminating methodological knowledge

Methodology Matters

Results reached with a **wrong procedure are unreliable**, though not necessarily wrong but they are open skepticism

Are we Getting Reliable Evidence?

Methodology is Critical in Empirical Studies

Natalia Juristo
Universidad Politécnica de Madrid

Project PID2022-137846NB-B-I00 funded by: