Industry Experiments

From the trenches

Anything that can go wrong will go wrong

Ayse Tosum
OULU

Burak Turham
OULU

Markku Oivo
OULU

Sira Vegas
UPMadrid

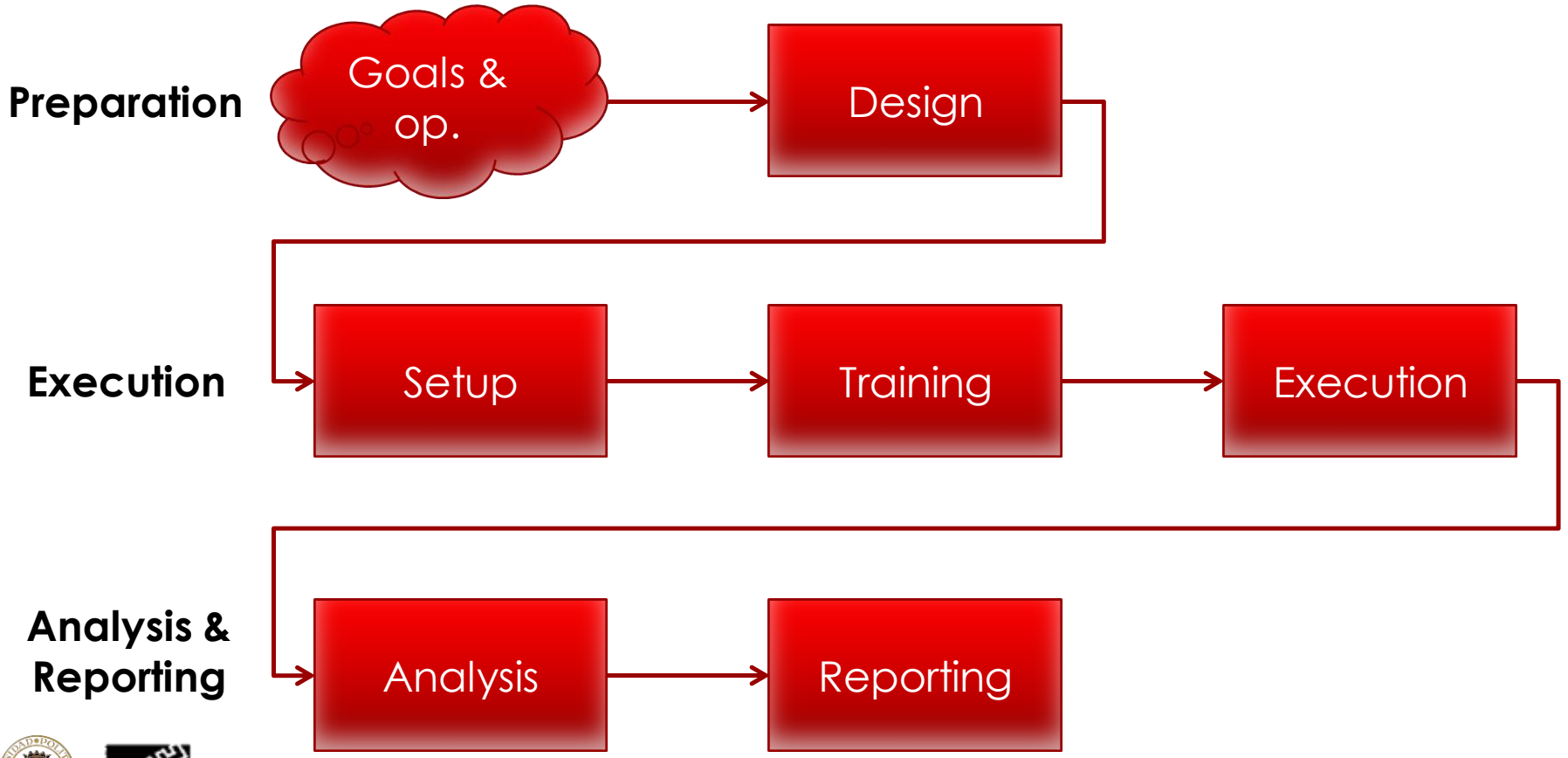
Natalia Juristo
UPMadrid & OULU

Davide Fucci
OULU

Oscar Dieste
UPMadrid

Hakan Erdogmus
CMU

Outline


Preparation
Goals & op.
Design
Execution
Setup
Training
Execution
Analysis & Reporting
Analysis
Reporting

Key point

- At a first sight, there is not a clear reason for differences between lab and field experiments
- However, there are, mostly due:
  - Psychology of the participants
  - Tighter restrictions imposed by the environment
  - Scarcity of resources
- Many of the problems may arise in lab experiment as well, but the exacerbate in industry
  - Lack of effective control
  - Curtailed improvisation

Real world vs. laboratory in SE

Real world	Lab
Professionals	Students
Real software	Toy software
Real projects	Exercises
Warm-up	Specific training

Design

- Design is concerned with ensuring control, e.g.: separation of effects, causality.
  - Typical example: factorial design
- However, control is not for free; it is a trade-off activity that depends heavily on:
  - Number of factors
  - Sample size
- Statistical power is the name of the game


Design

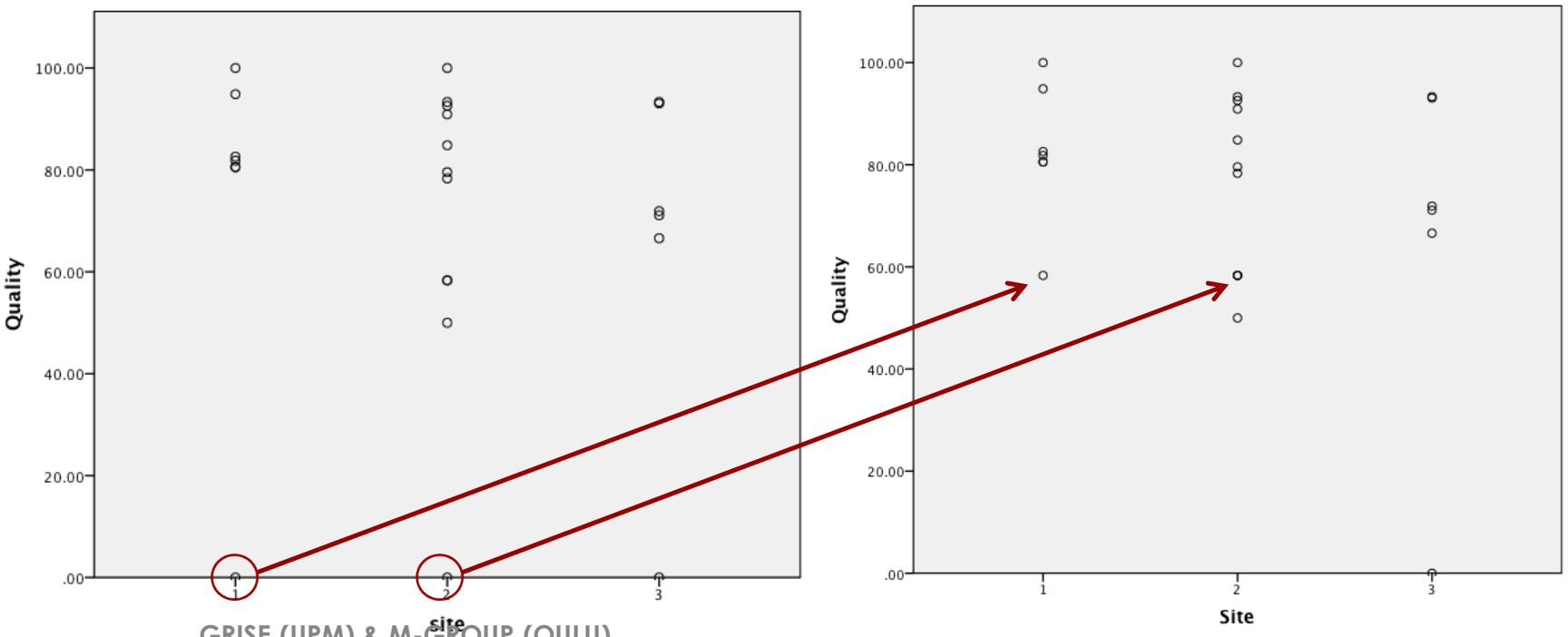
Power

- The number of experimental subjects (/units) in industry experiments tends to be (even)lower than in academic experiments.
  - In occasions much lower
  - Sample size has a strong effect on power
- Between-subjects designs usually do not achieve enough power.
  - Even fractional-factorial
- Where is the trouble?
  - If significant effects found, it may be due to small-size effects (or clerical errors, or…)
  - If not, results are inconclusive due to β


Design

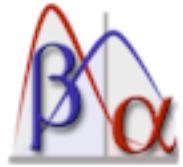
Power – illustrative example

- Are they too different?
  - Comparison become significant



Design

Power – concrete calculation



- Concrete calculation:
  - Assume (typical)
    - medium effect size (d = 0.5)
    - 2 treatments
    - α= 0.05
    - β= 0.80
    - 2 tails
  - Between subject design (based on t statistic):
    - 128 subjects required
  - Within-subjects design (paired t):
    - 34 subjects required

Design

Within-subjects designs

- Within-subjects designs are preferable, although they bring threats to validity to deal with
  - Maturity (learning, tiredness)
  - Period
  - Carry-over
- They can be managed
- We can fall into the temptation of using simpler designs (e.g.: AB) which confound treatment and task
  - Do not!
  - Crossover is a much better idea


Design

AB designs

- AB design seems the optimal quite often
- Typical example:

Group	Temp. sequence			
	Training	Exp. Session	Training	Exp. Session
G1	TLD	TLD	TDD	TDD
G2		TLD		TDD

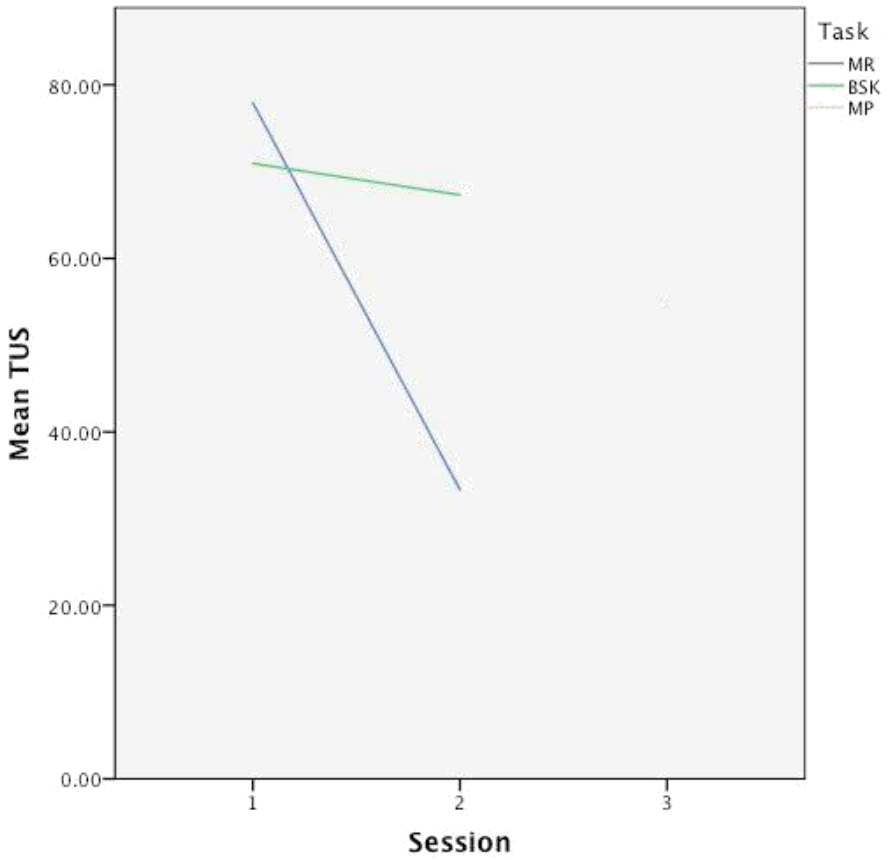
TABLE 4
Two-level, within-subjects design including the training course

- We are making an implicit assumption: there is not any *task * treatment* interaction

Design

AB design

- This plot shows the effect of
  - treatment (1 = ITL, 2 = TDD)
  - on task (MR, BSK)
  - in a company where crossover, instead of AB, was used
- *task * treatment* interaction is easily visible
- The same experiment gives opposite results depending on the assignment of task to session/treatment



Setup

- Setup is considered herein as the physical arrangement of everything needed to run the experiment
  - Facilities, including computers
  - Tools
  - Experimental Tasks
  - A little bit of everything, including documents, forms, etc.
- Sounds easy…


Setup

Facilities

- Depending on the particular experiment, some facilities should be reserved, computers configured, etc.
- In academic environments, classrooms and labs are readily available. In some cases, there are even supporting staff for those tasks
- In industry, comparable facilities may be, or not, available. In particular, computers are hard to obtain
  - Typical scenario: subjects bring their own *desktop* computers to a meeting room
  - Serious…


Setup

Tools

- Some experiments require specific instrumentation (e.g.: our subjects use a specific version of eclipse), or a given version of the JRE
- This may be easy or not depending on the company
  - Some preferred NOT to install anything in their computers (reasons: security, what else?)
  - Either case, it is impossible to check that the tools and related environment is installed correctly
    - In Java it may be possible, with C++ probably not
- We finally used virtual machines


Setup

Tools

- Virtual machines are fine, but:
  - Files are usually large
    - Take a couple big USB disk with you (see restrictions in next slide…)
  - The memory / processor in the host machine may not be large / powerful enough to run the guest system
    - Serious, again…
  - So simple things as the keyboard layout or USB support may create trouble


Setup

Tools

- Furthermore….
  - Companies may be quite picky with technology
  - In our case, the programming language (Java vs. C++) or unit testing framework (jUnit vs. gTests vs. Boost) were used in one or another experiment
- There are some implications:
  - Tasks (if legacy code or skeletons exist) must be ported
  - Related products (e.g.: test cases) should be ported as well
- Not only time consuming. It may also influence measurement (see later…)
  - e.g.: exception management in Java vs. C++


Setup

Restrictions

- Companies are concerned with security and confidentiality. They have policies that may create trouble with the experiment instrumentation
  - Network blocks some resources (e.g.: Dropbox, Github). It affects, e.g. the distribution of experimental materials or tools
  - Network services cannot be set up (e.g.: *syslogs*). Dynamic data collection may be impossible
  - Printing or storing data on servers depends on privileged accounts, not available all the time. It hardens data collection
  - Physical access to facilities may be impossible before/after a given time, or without supervision. It complicates setup checking or tuning

Setup

Tasks

- One of the strengths of industry experiments is the increased realism, in part due to the utilization of real software, in the scope of a real project, as tasks
- Natalia already explained that it was impossible
- But we are not that concerned (I?) about…

Setup

Tasks

- I am not sure that choosing a real tasks is a good idea
  - Hard to estimate complexity (e.g.: remember the AB design)
  - Domain knowledge effects (do developers do better due to treatment or task?)
    - Strong co-variable
  - Replication (disclosure and reutilization may be impossible due to confidentiality constraints)
- In my opinion, realism and validity do not necessarily match together


Real world vs. laboratory in SE

Less differences than expected…

Real world	Lab
Professionals	Students
~~Real software~~	Toy software
~~Real projects~~	Exercises
Warm-up	Specific training

Setup?

Subjects

- It has been mentioned before that assembling cohorts is complicated, and professionals may not have adequate commitment and motivation
- But professionals create further trouble:
  - Diversity: professionals tend to exhibit large differences in: education, experience, programming language, testing, design, tools, etc.

Much larger than students'

- Higher diversity means more noise, increased variance in the data

Training

- When subjects
  - do not know the technique or approach under test (the *treatment*) or
  - they have just little practice or
  - even when they are proficient, but they do not use the *treatment* in their daily work
- It is customary to provide some warm-up task or explicit training to
  - enable treatment application
  - and/or improve the similarity among subjects

Training

- Professionals are expected to be proficient
  - Only some warm-up required
- But they are not
  - Regular, full-fledged training necessary
- On the other hand, training, as Natalia said, is a good argument to convince companies to jump in
  - Life's hard

Real world vs. laboratory in SE

Before going on…

Real world	Lab
Professionals ??	Students
~~Real software~~	Toy software
~~Real projects~~	Exercises
Warm-up	Specific training

Subjects may be also uncommitted, non-motivated and diverse

Training

Provided

- Provide training is a source of problems in industry experiments:
  - Reluctance to training
  - Not-constructive discussion
  - Pressure on trainer
- Champion's participation may relieve those problems

Training

Provided

- Reluctance to training:
  - Subjects may exhibit a critical (*this course does not address my needs*) or hesitant (*this course won't teach me anything*) position
  - Experts tend to adopt the critical position. Due to prestige effects, they easily influence other subjects
  - These subjects' pre-conceptions give rise to other problems

Training

Provided

- Reluctance to training
- Not-constructive discussion:
  - Training is (/should be) ideologically neutral
  - Immoderate criticism lead to comparison with other approaches, trade-off evaluations, personal stories and opinions
    - The *treatment* may be regarded as ineffective by the subjects, paying thus less attention
    - Effective training time shortens

Training

Provided

- Reluctance to training
- Not-constructive discussion
- Pressure on trainer:
  - The trainer is asked or required to give an opinion about staff she may not know. It lessens the trainer's authority
  - When she knows, the explanations are usually time consuming. It is necessary to make double effort to keep the topic on track
  - Pretty exhausting

Training

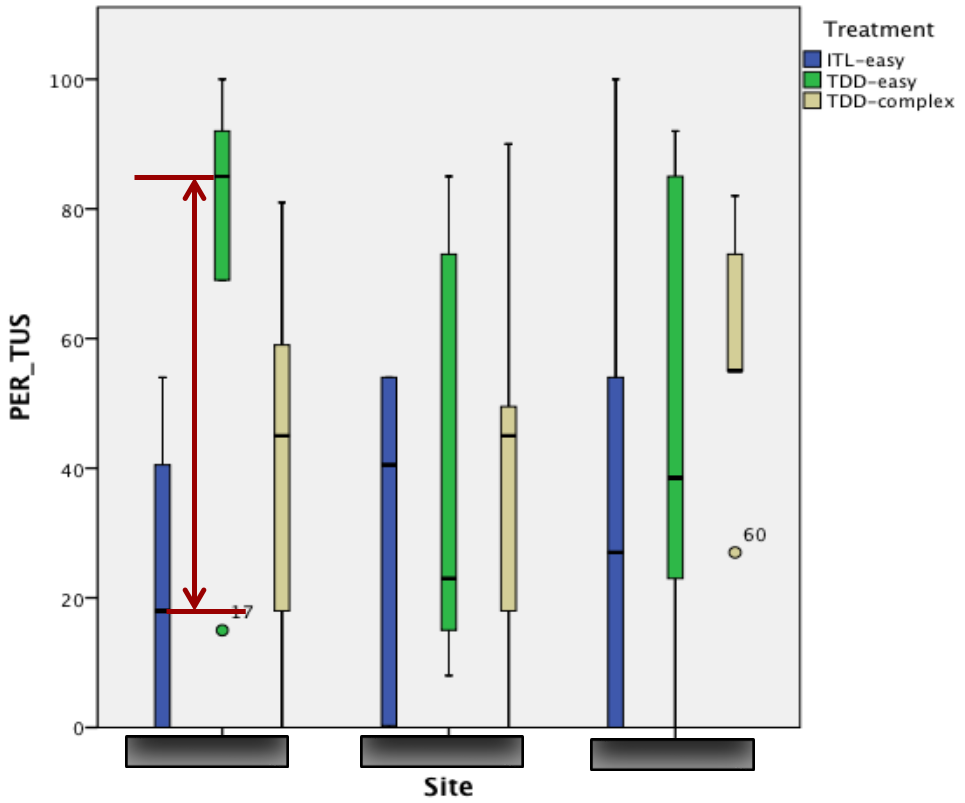
Provided

- Reluctance to training
- Not-constructive discussion
- Pressure on trainer
- When the champion is present, or when she participates in the training, subjects are more committed and/or relaxed
  - They are more committed when the champion has executive position
  - They are more relaxed when they recognize the champion as a competent professional in the domain


Training

Received

- Subjects' perception on training has a marked and almost immediate effect on motivation
  - Failure-to-train drops scores
  - Successful training raises scores above expected
- Trainer is a key element



Training

Received

- Some subjects asked for some kind of certificate or diploma that certifies that they attended the training
- They looked quite enthusiastic about that
- May some kind of certification could be used to improve commitment and motivation
  - Some kind of test could be proposed in this scenario, with the corresponding grade
  - Effective motivation in academic experiments? Why not in industry?


Execution

- During execution, subjects perform the experimental tasks
- No major trouble, with only two exceptions:
  - Interference
  - Lack of Conformance

Execution

Interference

- It takes many forms, but it is caused in all cases by the proximity between the regular working environment and the experimental environment
- You may expect:
  - Experimental materials are not available at the beginning or during the experimental session
  - Subjects may arrive late, leave early or be absent for prolonged times. They may even skip whole sections of the experiment
  - The experiment schedule needs to adapt to the working schedule (e.g.: meals, operation times, etc.)
    - Pretty rigid on this regard

Execution

Conformance

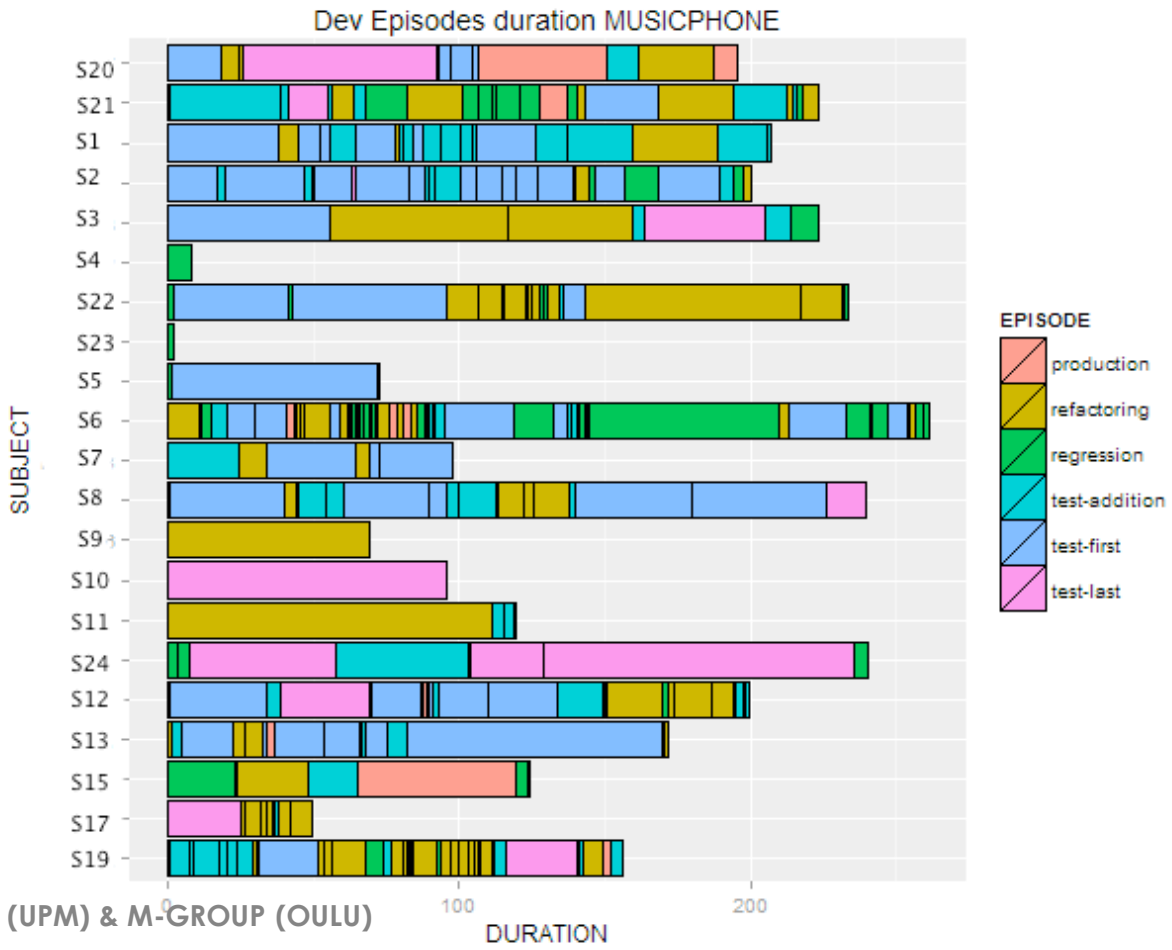
- Professional subjects tend to use their usual approach to solve the tasks, or revert quickly in case of trouble
  - Lack of conformance
  - What are we measuring?

Execution

Conformance

Execution

Conformance



Real world vs. laboratory in SE

Even worse…

Real world	Lab
Professionals ????	Students
~~Real software~~	Toy software
~~Real projects~~	Exercises
~~Warm-up~~	Specific training

Subjects may be also uncommitted, non-motivated, diverse and non-conformant !!

Analysis

- There are also some issues in a so neutral tasks as statistical analysis
  - Problems with measurements
  - Missing data
  - Large variability

Analysis

Measurement

- Raw data acquisition, in short
- In general, there are not differences with experiments in academy
- When different technologies are used, measurements are never alike
  - e.g.: constructors in Java vs. C++ (null checking)
  - e.g.: exceptions in Java vs. C++
- Depending on the measurement procedure, results may not be comparable on absolute scales
  - Use percentages


Analysis

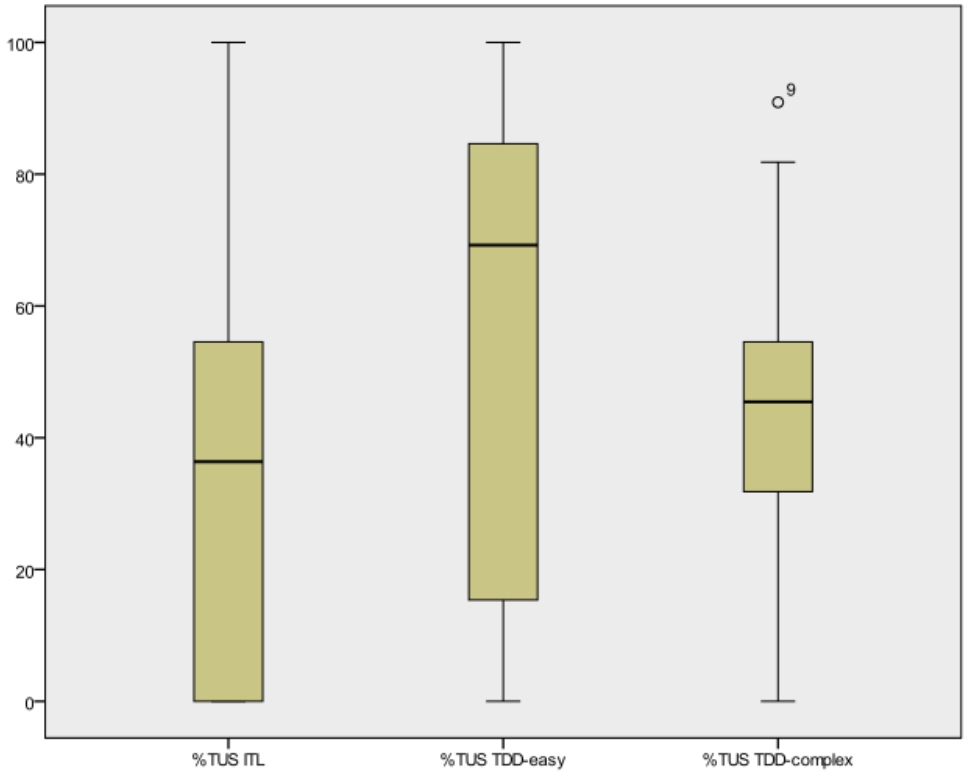
Missing data

- Dropouts or no-shows are very rare when using students
  - Typically they need to perform the experimental tasks to pass a course
- With professionals, there are often missing data
  - Interference, and also commitment
    - Subjects doing anything else during the experimental sesion
- Missing data renders data useless for some subjects
  - e.g.: 2 repeated measures


Analysis

Large variability in data

- The difference between low-performance subjects and high-performance subjects is very high
  - This could provoke that subjects have an influence on the results (apart from the treatment)
  - Again, an effect of interference, motivation and commitment
  - Conformance??
- It is even harder (remember: low power) to achieve statistical significance



Reporting

- Subjects who are reading the reporting do not have necessarily neither statistical nor experimental design background and/or knowledge
- This affects how the experiment has to be reported:
  - Simple descriptions
  - Simple plots
  - Meaningful effect size indexes


Analysis

Reporting

- The description of the experiment has to be done in a simple way
  - As simple as possible
- Try to avoid things which are important for researchers but not to practitioners, in particular
  - Justification of design decisions
  - Lengthy explanations of the statistical tests selected and why they have been chosen
  - Threats to validity are not treated separately, but incorporated to the discussion of results as possible limitations to them


Analysis

Reporting

- When reporting data analysis:
  - Simple, visual representation is critical. Histograms and boxplots are recommended
  - Check that the restrictions to apply the chosen statistical tests are met (e.g.: normality of sample/residuals, homoscedasticity, etc.), but do not bother them with these explanations
  - Base your recommendations not only in hypotheses testing (statistical significance), but also on effect sizes (practical significance) and power analysis.


Analysis

Reporting

- However, be careful with the measures chosen for effect size. Choose intuitive measures like:
  - Mean difference. Problem: unstandardized.
  - Common language (probability of superiority): "if Joe is taller than Sam, then there is a 63% probability that Joe's son is taller than Sam's son".
  - Eta and Partial eta squared: How much variability out of the total variability in the model the treatment explains.
  - Others:
    - Cohen's U3: "53 % of the treatment group will be above the mean of the control group"
    - Overlap: "92 % of the two treatment groups will overlap"
- Try to avoid non intuitive measures like Cohen's d

Summary

- Low Power
- Temptation of AB designs
- Facilities, including computers, not available
- Restrictions on tools, network, etc.
- Toy software
- Professionals are not a homogeneous population
- Resistance to training
- Training may influence motivation


Summary

- Interferences during execution
- Lack of conformance
- Non-trivial measurement, influenced by tools
- Drop-outs
- Large variability in data, making statistical significance harder to achieve
- Simple descriptions, clear pictures when reporting
- Intuitive effect size measures


Industry Experiments

Natalia Juristo (natalia@fi.upm.es)
Oscar Dieste(odieste@fi.upm.es)
http://grise.upm.es