UI/UX Advanced Lecture 2 (4)

How to set up experiments properly (Experimental design)

Alejandro Moreno
a.m.morenocelleri@saxion.nl

- Will be recorded and available in BB

Why do we run test?

Why do we run tests?

Find information

Why do we run tests?

Find out if something works
Find out what to fix
Find out if something is enjoyable
...

Research methods

Desk research

User testing
- Give a task
- Scenario/Exploration

Interviews / Questionnaires / Surveys

(Non)Participatory observation

Research methods

Desk research

User testing
- Give a task
- Scenario/Exploration

Interviews / Questionnaires / Surveys (Next week)

(Non)Participatory observation

Research methods

Desk research

User testing

- Give a task (Today)
- Scenario/Exploration

Interviews / Questionnaires / Surveys

(Non)Participatory observation

How do we set up tests/experiments properly?

Experimental Design

Test vs Experiment

Setting up

Experiments

What is the purpose of an experiment?

Validate a hypothesis

You want to find out whether your hypothesis is true or not

A failed experiment is one that was not designed properly – if you find out that you were “wrong”, that is still valid knowledge

Hypothesis

Design

Analysis of results

Running an experiment requires careful planning in advance

Participants

Apparatus (means)

Procedure

Design

Experimental design

4 Considerations

Participants

People who will test our product

How many do we need? – *Sample size*

- More is better
  - Depends on margin of error and confidence interval (95%)

How do you recruit/select them? - *Sampling*

- Convenience – we take what we can get
- Purposeful – we have specific criteria
- Snowball – participants recruit other participants

Apparatus

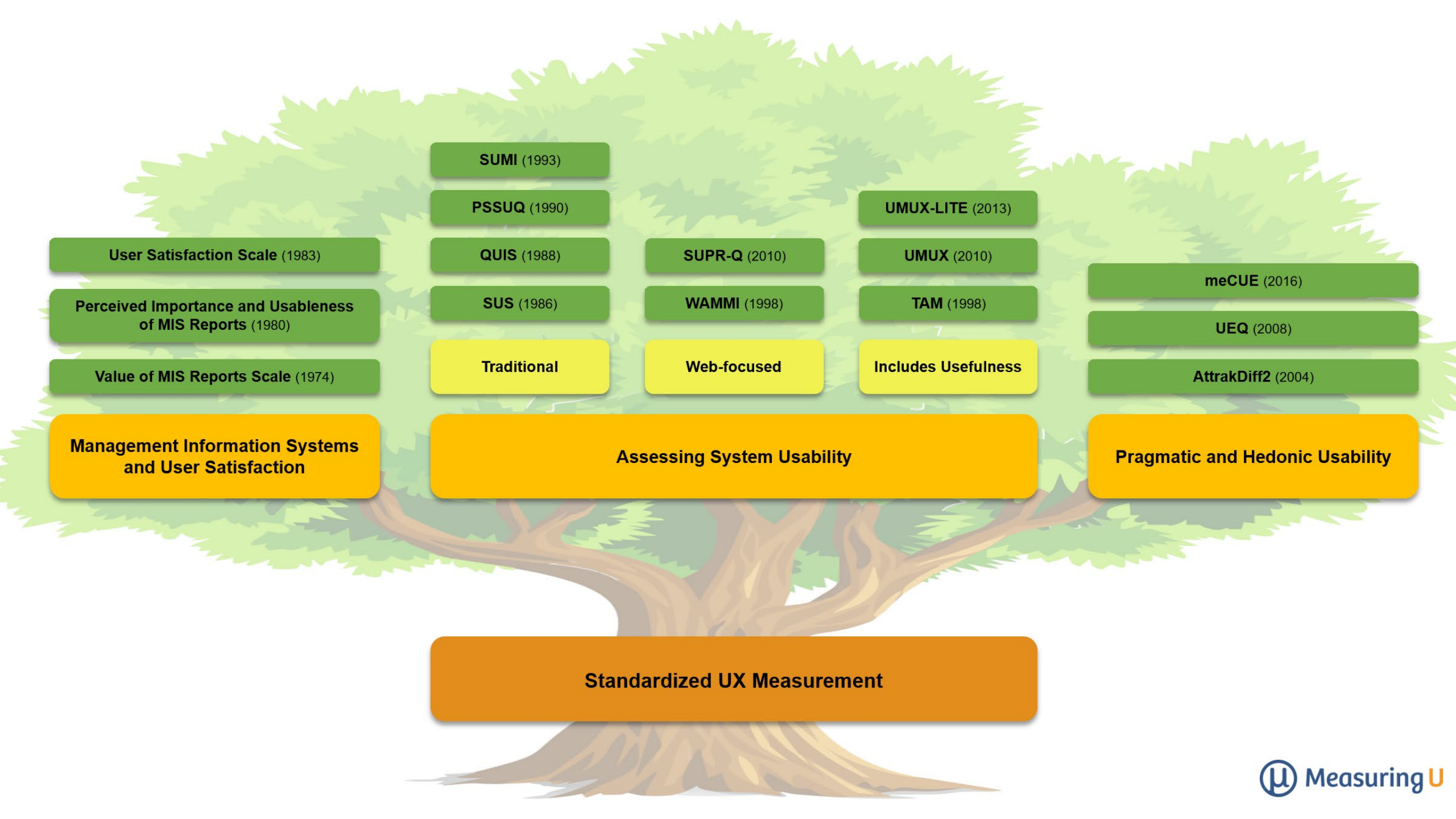
What we need in terms of equipment, space, data, etc.

Physical objects but also digital tools

- Cameras
- Rooms
- Software
- …

Validated questionnaires

- System Usability Scale (SUS)
- Usability Metric for User Experience (UMUX)
- Technology Acceptance Model (TAM)
- Game Experience Questionnaire (GEQ)


SUMI (1993)
PSSUQ (1990)
QUIS (1988)
SUS (1986)
UMUX-LITE (2013)
UMUX (2010)
TAM (1998)
User Satisfaction Scale (1983)
Perceived Importance and Usableness of MIS Reports (1980)
Value of MIS Reports Scale (1974)
SUPR-Q (2010)
WAMMI (1998)
meCUE (2016)
UEQ (2008)
AttrakDiff2 (2004)
Traditional
Web-focused
Includes Usefulness
Management Information Systems and User Satisfaction
Assessing System Usability
Pragmatic and Hedonic Usability
Standardized UX Measurement

Procedure

What will participants do?

Changes in procedure may affect the validity of the results

- Defining a protocol/plan is important

Your participants are human beings!

- Welcome them and explain what they will do
  - Without giving away your goal
- Have a consent form signed
- Give controlled breaks
- Be kind and grateful
- Debrief them at the end (and thank them)
  - Give them a cookie

Design

What are we trying to validate, and how to best get this done

Variables

What we manipulate		What we measure
Independent variable	→	**Dependent variable(s)**

Example

How would the difficulty of a test change based on the type of question I have?

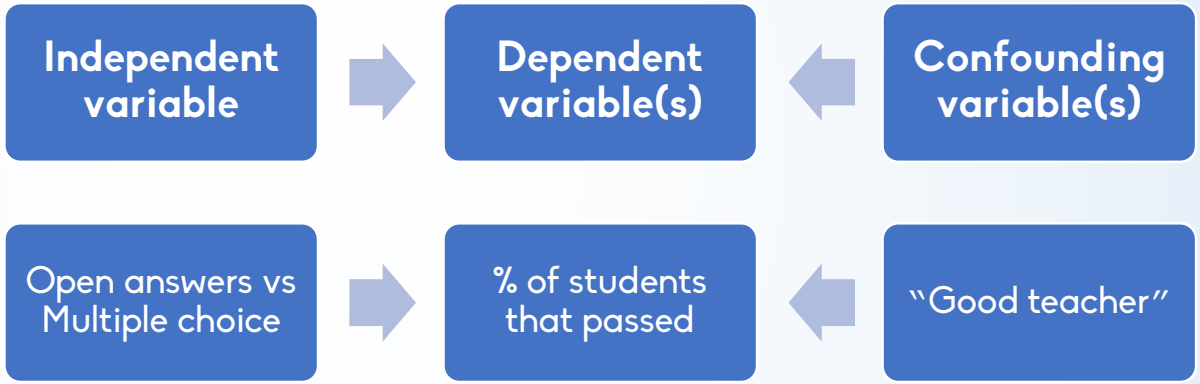
Independent variable		**Dependent variable(s)**
Open answers vs Multiple choice	→	% of students that passed

Design

What are we trying to validate, and how to best get this done

Common mistake - Confounding variables

Variables that introduce uncontrolled variation



Design

What are we trying to validate, and how to best get this done

Variables

What we manipulate

Independent variable

What we measure

Dependent variable(s)

Lets try this out

Scenario: I am interested in changing my personal logo, but I don't know what would be the best choice

Design

What are we trying to validate, and how to best get this done

Variables

What we manipulate

Independent variable

What we measure

Dependent variable(s)

Lets try this out

Scenario: I am interested in changing my personal logo, but I don't know what would be the best choice

Design

What are we trying to validate, and how to best get this done

Variables

What we manipulate

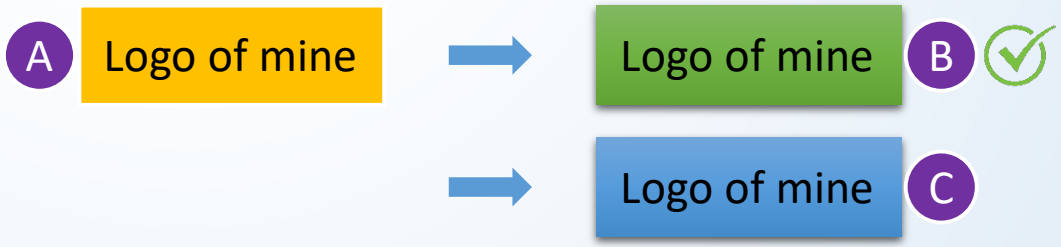
Independent variable

What we measure

Dependent variable(s)

Lets try this out

Scenario: I am interested in changing my personal logo, but I don't know what would be the best choice



Design

What are we trying to validate, and how to best get this done

Variables

What we manipulate

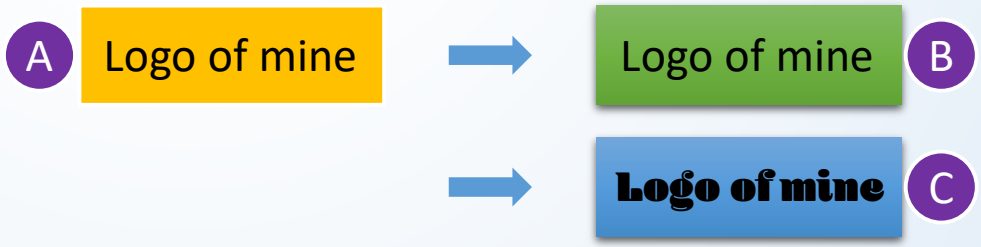
Independent variable

What we measure

Dependent variable(s)

Lets try this out

Scenario: I am interested in changing my personal logo, but I don't know what would be the best choice



Design

What are we trying to validate, and how to best get this done

Variables

What we manipulate

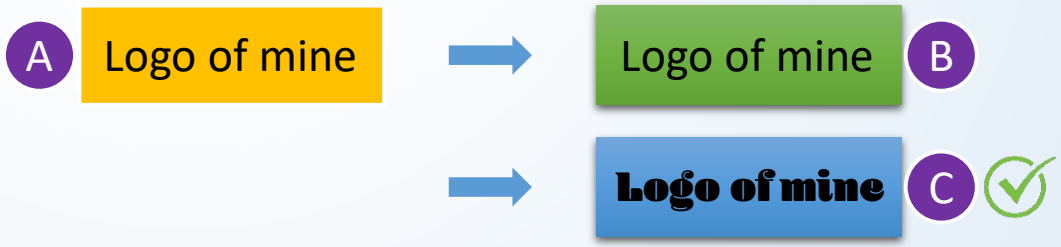
Independent variable

What we measure

Dependent variable(s)

Lets try this out

Scenario: I am interested in changing my personal logo, but I don't know what would be the best choice



Design

What are we trying to validate, and how to best get this done

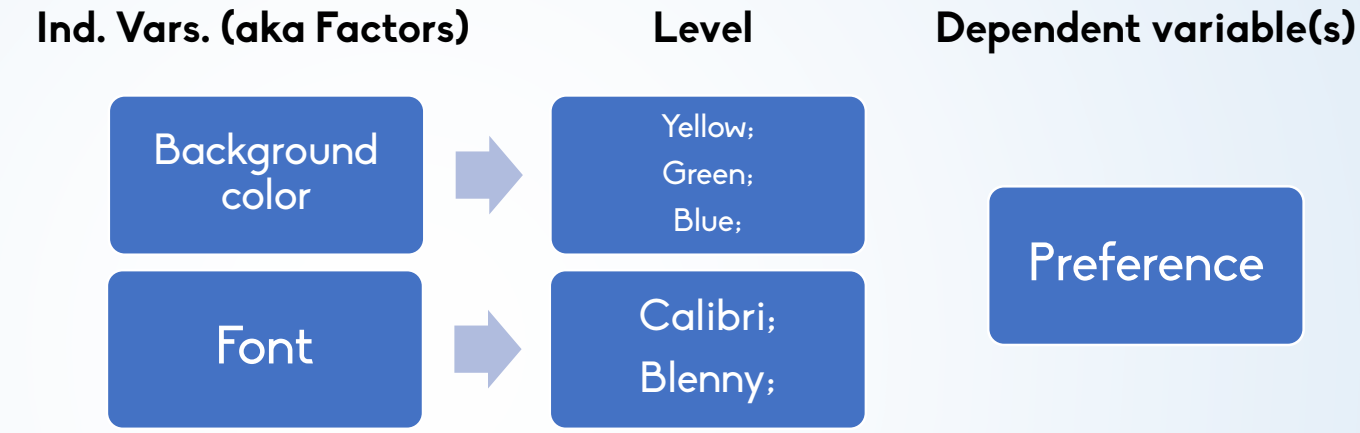
Factorial design

One Factor Design

Design

What are we trying to validate, and how to best get this done

Factorial design



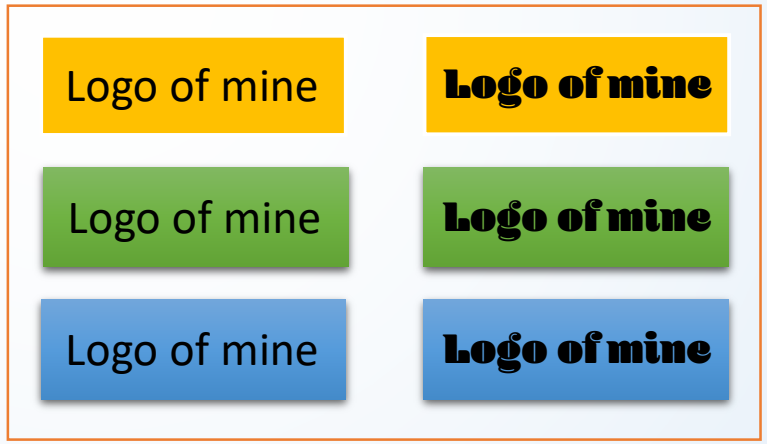
3 x 2 Factorial Design = 6 conditions

Background color **Font**

Design

What are we trying to validate, and how to best get this done

3 x 2 Factorial Design = 6 conditions

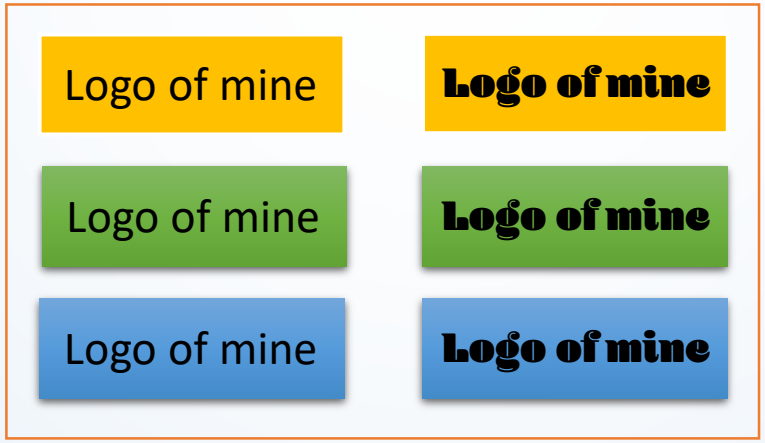


3x2 factorial design

Design

What are we trying to validate, and how to best get this done

Common Mistake – Incomplete factorial design



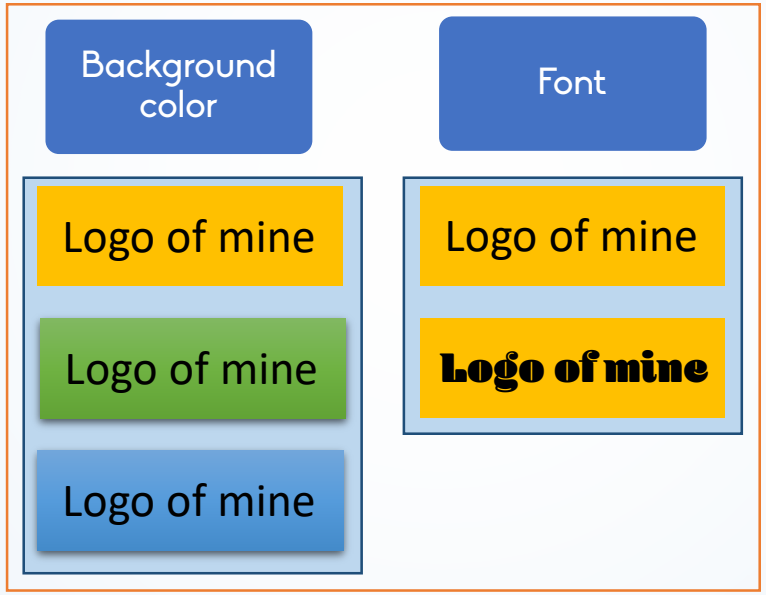
3x2 factorial design

Incomplete 3x2 factorial design

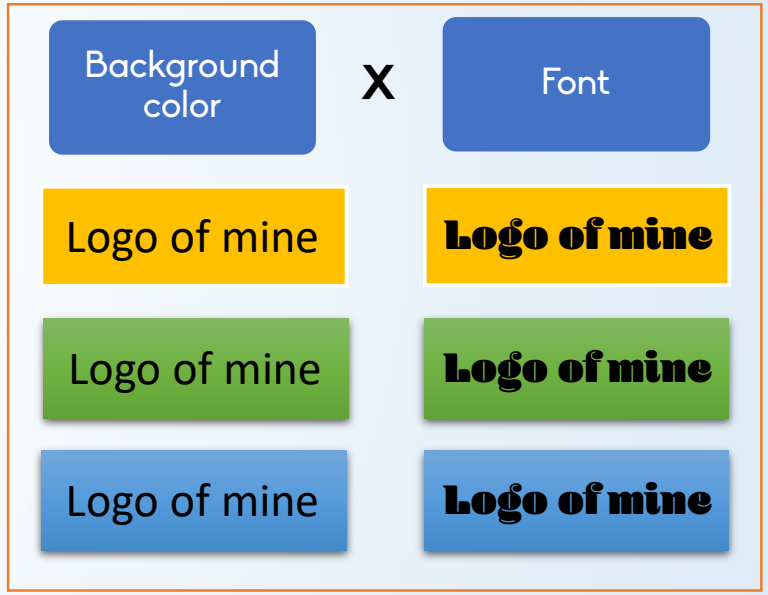
Design

What are we trying to validate, and how to best get this done

The main advantage of Multifactorial Design is the possibility of **understanding the effect that factors have together and on each other**



2 x One-factor design

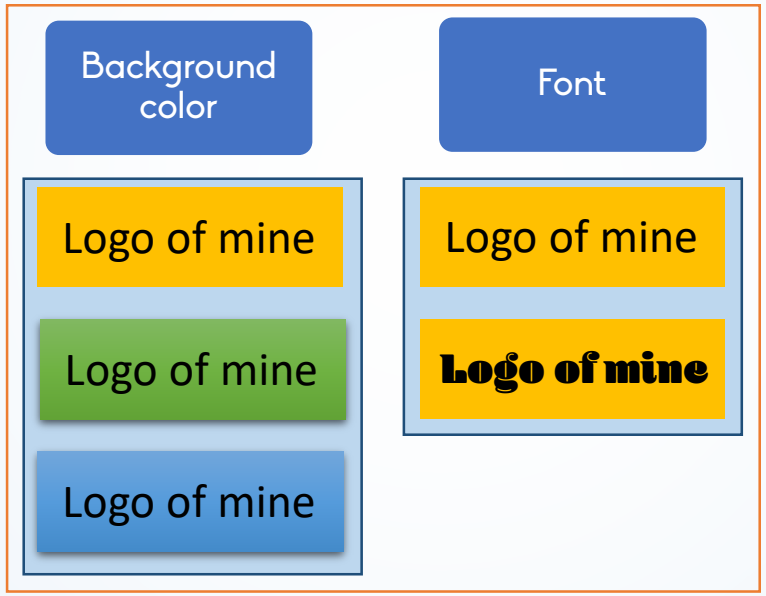


3x2 factorial design

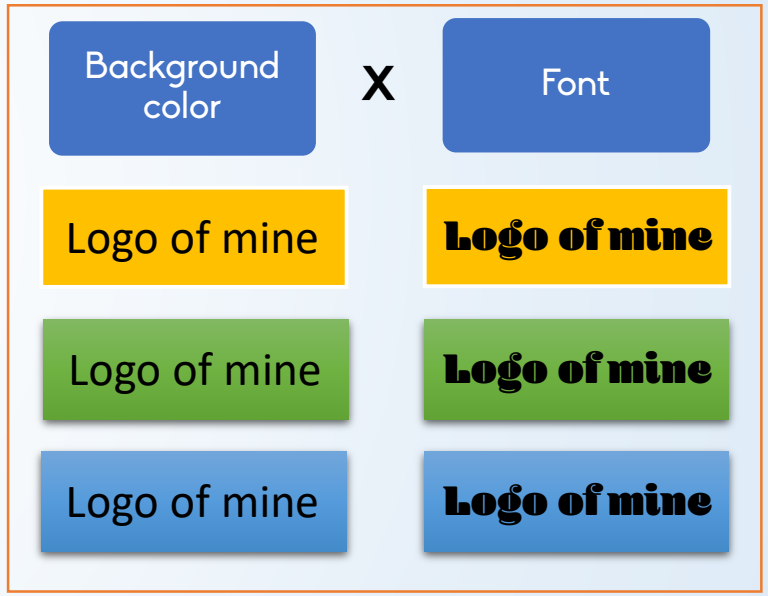
Design

What are we trying to validate, and how to best get this done

The main advantage of Multifactorial Design is the possibility of **understanding the effect that factors have together and on each other**



2 x One-factor design



3x2 factorial design

Try to read this!

Design

What are we trying to validate, and how to best get this done

Between and within-subjects design

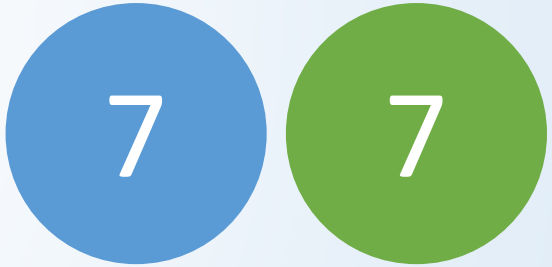
Between-subjects
One participant tests only one condition

A B 8

P1 P3 P5 P7 | P2 P4 P6 P8

Within-subjects
Each participant tests all conditions

A B 7



P1-7 | P1-7

Design

What are we trying to validate, and how to best get this done

Within-subjects
Each participant tests all conditions

Protocol
1
2
3
4
5
6

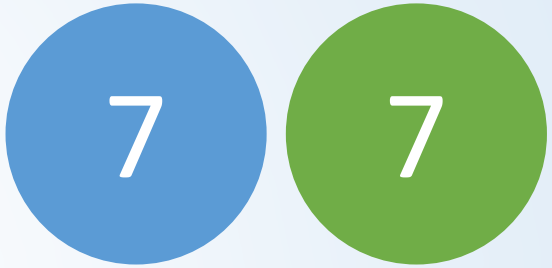
Design

What are we trying to validate, and how to best get this done

Common Mistake – Order effect

Within-subjects
Each participant tests all conditions

A B 7



Participant	1st condition	2nd condition
Alejandro	A	B
Yvens	A	B
Boris	A	B
…	A	B

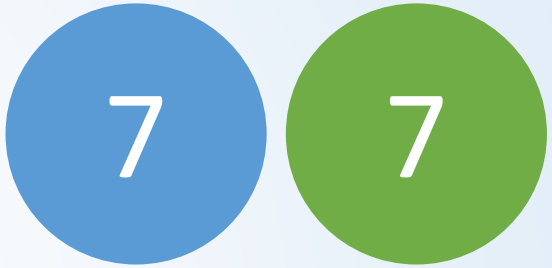
Would your results change if you did B and then A?
Fatigue, Practice, Learning effects

Design

What are we trying to validate, and how to best get this done

Common Mistake – Order effect

Within-subjects
Each participant tests all conditions

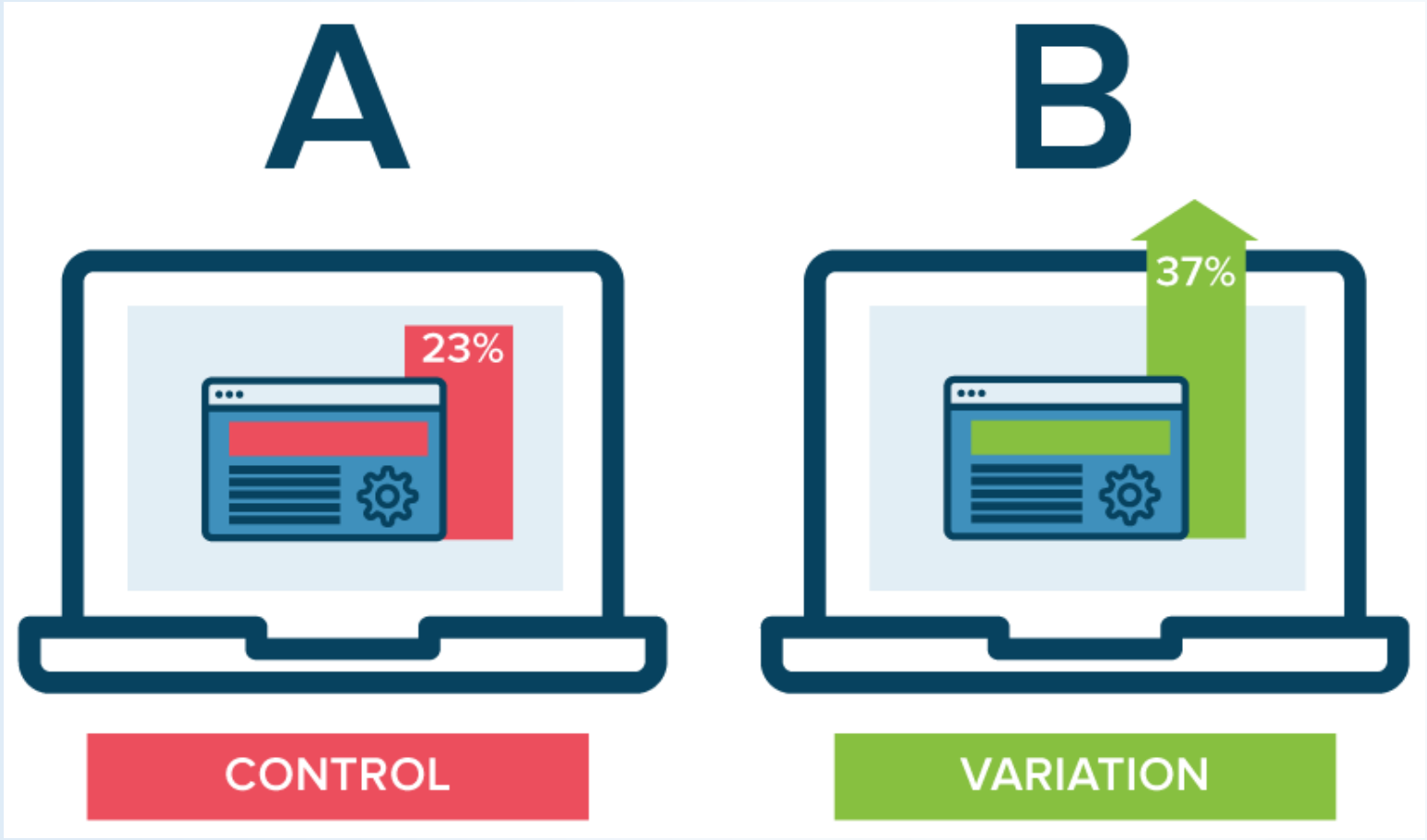


Participant	1st condition	2nd condition
Alejandro	A	B
Yvens	B	A
Boris	A	B
…	Alternate conditions (AB, BA, AB, …)	

Counterbalancing is needed


I'M A/B TESTING TO SEE WHAT GETS THE MOST CANDY.
TOM FISH BURNE
© marketoonist.com

A/B Testing



A/B Testing

Two (or more) versions of the product

Some users are exposed to one version

You design the variations to test specific changes (hypothesis)

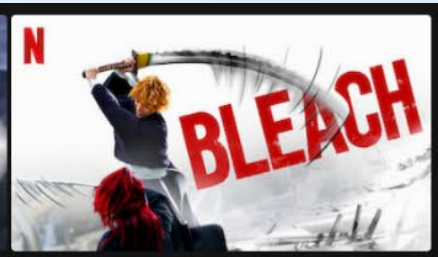

IT'S KIND OF A FUNNY STORY


CROUCHING TIGER, HIDDEN DRAGON
SWORD OF DESTINY


THE DO OVER


KNIGHTFALL


BERSERK


BLEACH


BERSERK


FULLMETAL ALCHEMIST


BAD BOYS FOR LIFE


CASTLEVANIA


GREAT PRETENDER


THE DRAGON PRINCE

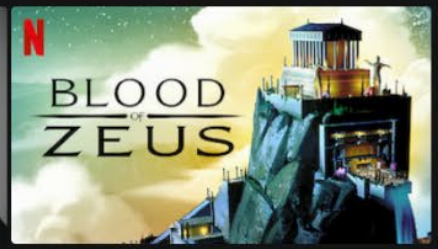

WARRIOR NUN


Cagaster of an Insect Cage


JESSICA JONES


KENGAN ASHURA


BBC
LUTHER


BLOOD OF ZEUS


HI SCORE GIRL


SPACE FORCE


DRAGON'S DOGMA


Sirius the Jaeger


Teasing Master Takagi-san


SEIS MANOS


BERSERK


CROUCHING TIGER, HIDDEN DRAGON
SWORD OF DESTINY


BLEACH


BAD BOYS FOR LIFE


KILLER ELITE


IT'S KIND OF A FUNNY STORY


KNIGHTFALL


FULLMETAL ALCHEMIST


CASTLEVANIA


GREAT PRETENDER


WARRIOR NUN


THE DRAGON PRINCE


JESSICA JONES


Eagaster of an Insect Cage


KENGAN ASHURA


SPACE FORCE


BBC
LUTHER


HI SCORE GIRL


DRAGON'S DOGMA


Sirius the Jaeger


BLOOD OF ZEUS


Teasing Master Takagi-san


SEIS MANOS

https://netflixtechblog.com/selecting-the-best-artwork-for-videos-through-a-b-testing-f6155c4595f6

Cells	Cell 1 (Control)	Cell 2	Cell 3
Box Art	 Default artwork	 14% better take rate	 6% better take rate

Common A/B testing procedure

Define hypothesis

Define independent variables (preferably 1)

Define metrics (dependent variables)

Create a "control" and a "challenger"

Decide on design (within or between)

Split participants equally amongst all conditions

Decide on sample size and statistical significance

Test variations simultaneously

Conduct statistical analysis on your metrics

Take home points!

- Experiments are carried out to validate hypotheses.
- When designing an experiment, we need to consider Participants, Apparatus, Design, and Procedure.
- Independent variables are the aspects of an experiment we are in control of – dependent variables are the metrics.
- One factor design provides reliable results (cause-effect), but multifactorial design can provide better insights by sacrificing simplicity.

Take home points!

- Between-subjects design is preferred, but requires a high number of participants. Within-subjects design requires less participants, but is more prone to order effects (fatigue, learning, practice).

- Common mistakes in A/B testing
  - Testing too many things simultaneously (too many factors makes it hard to understand cause-effect)
  - Not testing all conditions (incomplete factorial design)
  - Unbalanced participant numbers
  - Confounding variables
  - Order effects

Thanks!

Alejandro Moreno
a.m.morenocelleri@saxion.nl