

Experimental design and analysis

Introduction to experimental design

<https://www.lri.fr/~appert/eval/>

TODO before next week

We will use Jupyter notebook for statistical analyses.

You have to install ANACONDA.NAVIGATOR

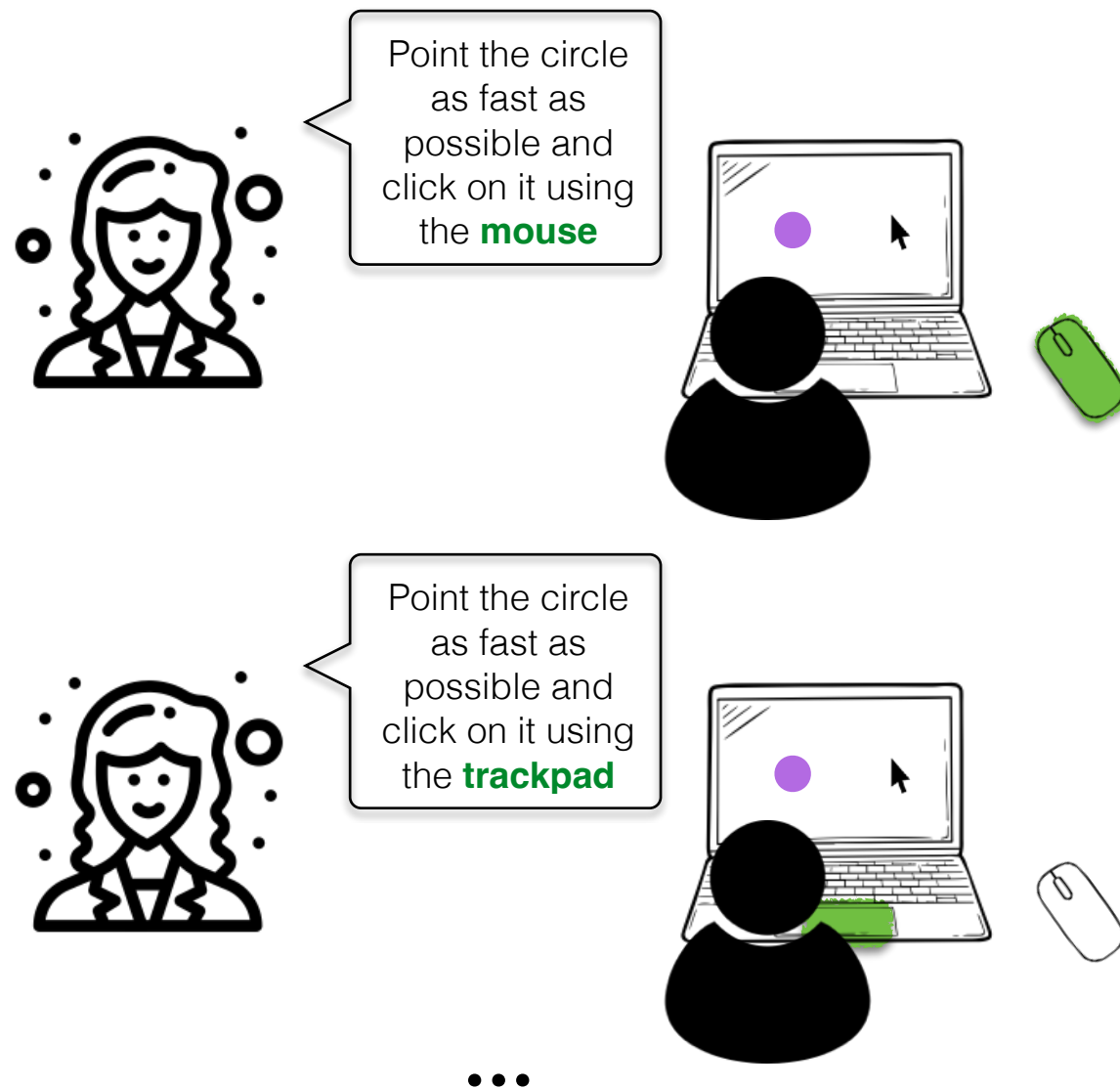
<https://docs.anaconda.com/anaconda/install/>



Hypothesis

Laboratory experiment - overview

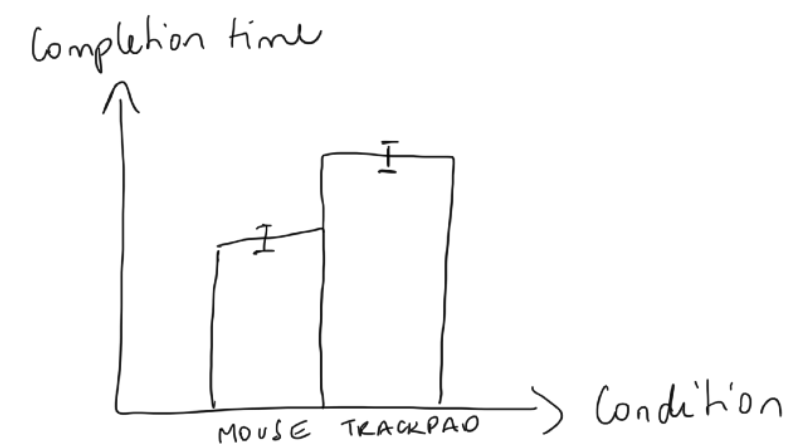
Hypothesis: Users point faster with a mouse than with a trackpad



1. The operator asks participants to complete tasks under **specific conditions**

| Participant | Condition | Completion time |
|-------------|-----------|-----------------|
| P1 | mouse | 403 |
| P1 | trackpad | 527 |
| ⋮ | ⋮ | ⋮ |
| P2 | mouse | 522 |
| P2 | trackpad | 608 |
| ⋮ | ⋮ | ⋮ |

2. Participants' performance is recorded in log files



3. Log files are analyzed with statistical procedures to test the research hypothesis

More formally, a laboratory experiment is...

...a test that is made to demonstrate a known truth, [examine the validity of a hypothesis](#), or determine the efficacy of something previously untried.

It is conducted under highly controlled conditions (not necessarily a laboratory), where accurate measurements are possible. The researcher decides where the experiment will take place, at what time, with which participants, in what circumstances and using a standardized procedure.

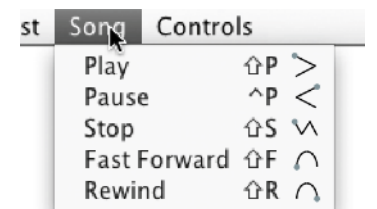
Research Hypothesis

A laboratory experiment starts with a [research hypothesis](#)

H₁: Users point faster with a mouse than with a trackpad



H₂: Gesture commands are easier to recall than keyboard shortcuts



H₃: Users make more typing errors with software keyboards than with physical keyboards



...

What is a hypothesis?

A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation

A hypothesis should be:

testable: the means for manipulating the variables and/or measuring the outcome variable must exist

falsifiable: must be able to disprove the hypothesis with data

precise: should be specific (**operationalized**)



Very important

Testing a hypothesis

The experimenter
manipulates **factor(s)** (aka. independent variables)
and collects **measure(s)** (aka. dependent variables)

H: Users point faster with a mouse than with a trackpad

what is manipulated
(experimenter sets the value)

Factor(s)

e.g., Pointing Device
∈ {Mouse, Trackpad}

what is measured
(depends on participants'
performance and preferences)

relationship

experimental task
(e.g., pointing a target)

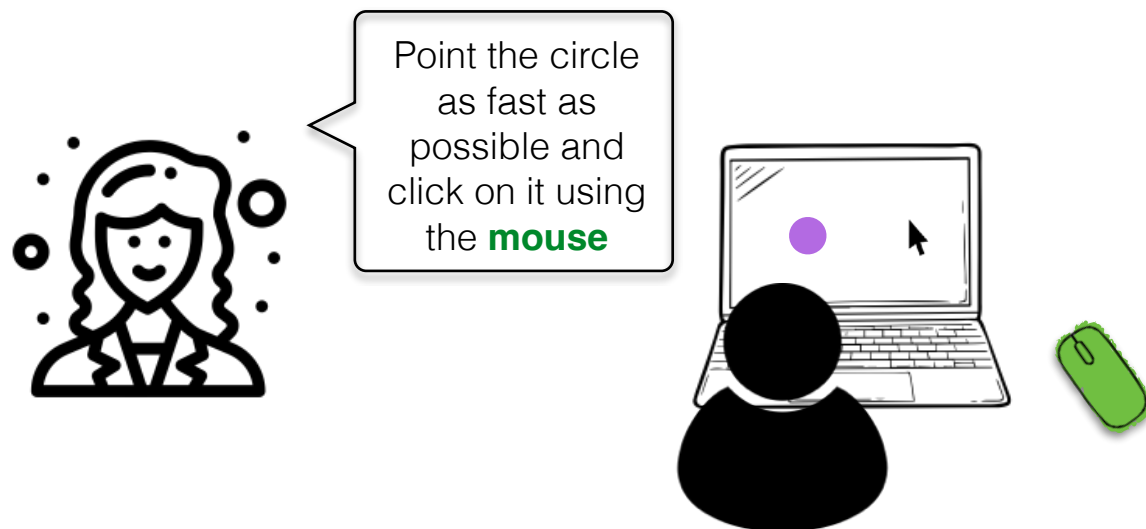
Measure(s)

e.g., Completion time

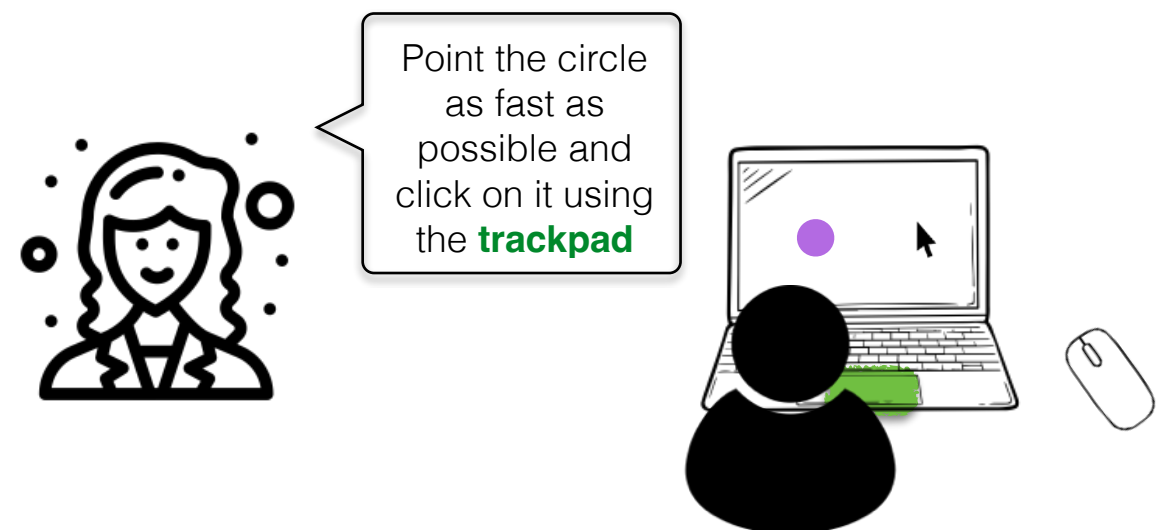
Testing a hypothesis

The experimenter
manipulates **factor(s)** (aka. independent variables)
and collects **measure(s)** (aka. dependent variables)

The experimenter **decides on** the value
of Factor *pointing device*
pointing device = mouse



The experimenter **decides on** the value
of Factor *pointing device*
pointing device = trackpad



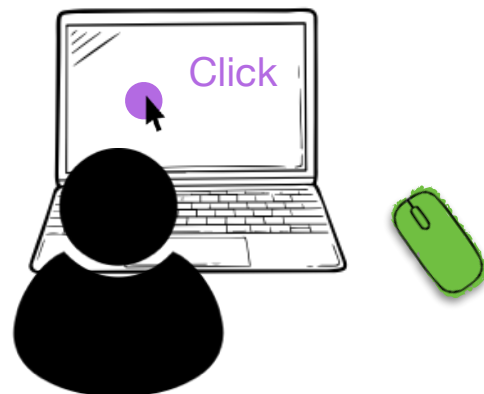
Testing a hypothesis

The experimenter

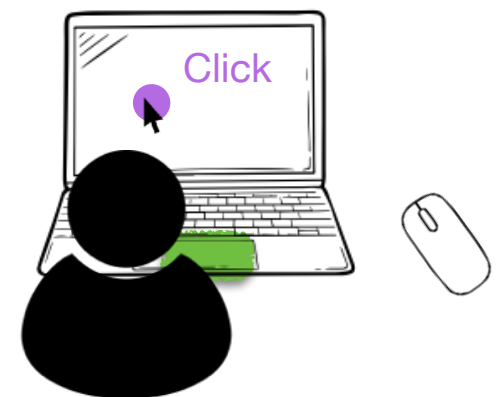
manipulates **factor(s)** (aka. independent variables)

and collects **measure(s)** (aka. dependent variables)

The experimenter **observes/collects**
the value of Measure *completion time*
completion time = 425ms



The experimenter **observes/collects**
the value of Measure *completion time*
completion time = 512ms



Falsifying a hypothesis

A hypothesis makes a **general statement**

An experiment collects a **specific data sample**

Two cases:

1. If the sample is coherent with the hypothesis, you cannot validate the hypothesis (another sample may have been different and inconsistent with the hypothesis)
2. If the sample contradicts the hypothesis, you have identified a counter-example so the hypothesis cannot be true.

Reasoning with the null hypothesis

The experimenter's research hypothesis H expects a difference between two conditions.

H : There is a difference between $C1$ and $C2$.

The null hypothesis is

H_0 : There is no difference between $C1$ and $C2$.

If the collected sample during the experiment reveals a difference between $C1$ and $C2$, we can reject H_0 .

The experimenter can conclude that there is a difference in this specific experiment, thus **supporting (not validating!)** H .

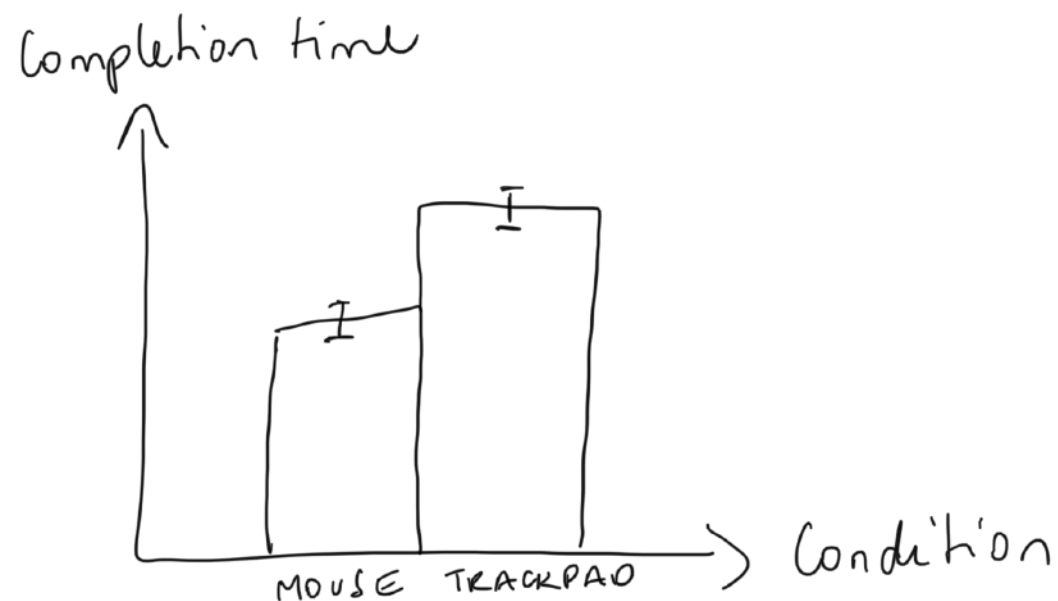
NB: *Supporting* is not as strong as *Validating*. Be careful not to claim that you have validated a hypothesis with an experiment, you can just support it.

Testing a hypothesis



research hypothesis: Users point faster with a mouse than with a trackpad

null hypothesis: Users point as fast with a mouse as they point with a trackpad



*According to collected data,
participants are significantly faster
with the mouse than with the trackpad*

=> this rejects the null hypothesis

*It does not validate the research
hypothesis, but **supports** it*

Task design

Operationalization

Operationalization defines a fuzzy concept so as to make it distinguishable, measurable, and understandable by empirical observation.

“Children grow more quickly if they eat vegetables”

Operationalizing entails defining terms:

'children' = $4 < \text{age} < 8$

'vegetables' = quantity of vitamin C

'Grow more quickly' = cm per year

Operationalization and laboratory experiments

Operationalizing a hypothesis to test in a laboratory experiment means identifying three things:

Factors

Measures

An experimental task that turns measures into a function of factors

The task is designed so that If I observe a change in a **measure** (e.g., completion time), it is because of a change in a **factor** (e.g., pointing device).

Confounding variable (bias)

Operationalizing often entails to simplify a task to its minimum so as to eliminate bias and effects from confounding variables. A confounding variable is any variable other than the factor that can possibly explain the change in measures.

Learning can be a confounding variable

e.g. all participants are first tested with the physical keyboard and then with the software keyboard → software keyboard has the advantage that participants have learned the keyboard layout

Prior experience can be a confounding variable

e.g. use conventional keyboard shortcuts (e.g. ctrl+V for paste) when comparing them to gesture shortcuts, which are a non-familiar type of shortcuts → keyboard shortcuts are favored because of participants' prior knowledge

Validity issues

Operationalizing often entails simplifying the phenomenon of interest. It requires to find the good trade-off between **internal validity** and **external validity**.

Internal validity

The experiment is sound so that observed effects are actually attributable to the manipulated factors.

External validity

The experiment is not too simplistic so results can generalize to other subjects and situations.

Internal validity

Causality and Correlation

An experiment is internally valid only if there is a **causal relationship** between factors and measures.

Correlation

mathematical relationship between two variables.

Causality

physical relationship between two variables. There is a chain of events when the first variable varies that causes the other variable to vary (involves time).

Internal validity

Causality and Correlation

Correlation does not imply causality

For example, we noted a high correlation between the weight and height of persons.

However, high weight \Rightarrow high height is not true!

The problem is that mathematics cannot distinguish correlation from causality. When can we say that correlation imply causality?

When the experiment design is done with appropriate care to avoid confounding and other threats to the **internal validity** of the experiment.

The task is designed so that If I observe a change in a **measure** (e.g., completion time), it is because of a change in a **factor** (e.g., pointing device).

Operationalization - task design

Well-known standards

If some well-known standards exist, use them

Pointing

ISO. 9241-9 Ergonomic requirements for office work with visual display terminals (VDTs)-Part 9: Requirements for non-keyboard input devices.

Text entry

MackKenzie et al.'s phrase set

<http://www.yorku.ca/mack/chi03b.html>

Operationalization - task design

Well-known standards - Pointing

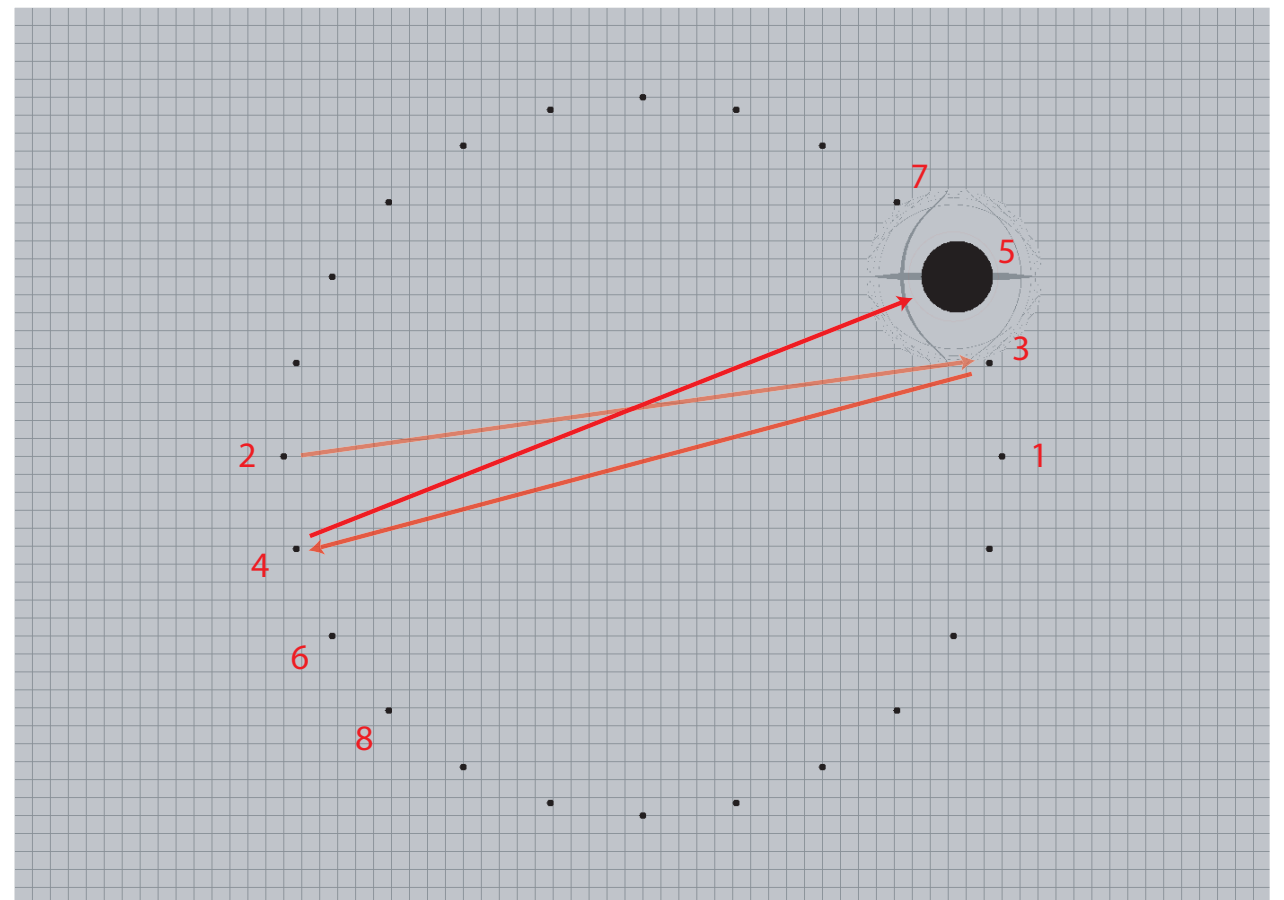
H: Users point faster with a mouse than with a trackpad

Factors: input device, pointing difficulty

Measures: completion time

Task: point the target as fast and as accurately as possible.

The experiment design should use a circular layout and specific order of appearance of successive targets to force participant to point in every direction (9241-9 ISO standard).



Operationalization - task design

Well-known standards - Text entry

H: Users are more accurate with physical keyboards
than with software keyboards

Factor: keyboard type

Measures: typing speed,
typing errors

Task: copy the sentence as
fast as and as accurately
as possible.

The experiment design should use a
a phrase set that is representative of
the target language.

video camera with a zoom lens
have a good weekend
what a monkey sees a monkey will do
that is very unfortunate
the back yard of our house
I can see the rings on Saturn
this is a very good idea
...



Excerpt of MacKenzie & Soukoreff's phrase
set (2003). It is a set of phrases that are
moderate in length, easy to remember, and
that have digram frequencies that are
representative of English.

Operationalization

First, look at what others have done^(*) when they have tested hypotheses that are similar to your research hypotheses.

If you cannot find good examples... Think carefully to define an experiment that ensures a good trade-off between internal and external validity

The task is designed so that If I observe a change in a **measure**, it is because of a change in a **factor**.

Internal

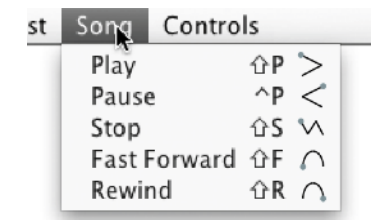
The experiment design is representative of actual use.

External

^(*) In trustable sources like standards and academic articles that were published in peer-reviewed journals / conferences.

Operationalization

No standard defined - Example#1



H: Gesture shortcuts are easier to learn than Keyboard shortcuts

Factor: Type of shortcut {Gesture, Keyboard}

Measure: Recall rate

Task: Ask the participant to perform the right shortcut in response to a command stimulus.

Recall score is 1 if participant performs the right shortcut, 0 otherwise.

Operationalization

No standard defined - Example#1



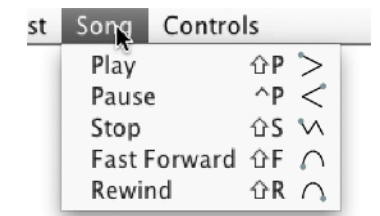
Example of questions to assess the validity of our experiment:

Should we use existing shortcuts (like Cmd+C for copy command, Cmd+V for paste command)?

How many shortcuts should we consider?

Operationalization

No standard defined - Example#1

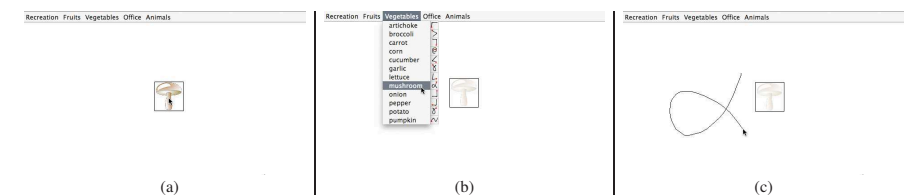


Example of questions to assess the validity of our experiment:

Should we use existing shortcuts (like Cmd+C for copy command, Cmd+V for paste command)?

No because of prior experience that we cannot easily control. For a better internal validity, let's rather use arbitrary mappings for both types of shortcuts.

| ICON | Keys | Stroke | ICON | Keys | Stroke |
|------|---------|--------|------|--------|--------|
| | Shift+W | | | Ctrl+W | |
| | Shift+D | | | Ctrl+D | |
| | Shift+Q | | | Ctrl+Q | |
| | Shift+S | | | Ctrl+S | |
| | Shift+E | | | Ctrl+E | |
| | Shift+A | | | Ctrl+A | |
| | Shift+R | | | Ctrl+R | |



How many shortcuts should we consider?

The literature informs us that 14 is representative of expert usage. Our findings should thus transfer to real context of use (external validity)

Operationalization

No standard defined - Example#1



Example of questions to assess the validity of our experiment:

Is a use of shortcuts right after learning them is representative of a real context of use?

Operationalization



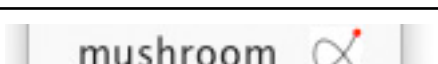
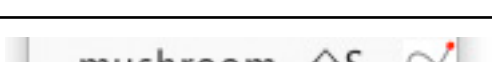
No standard defined - Example#1



Example of questions to assess the validity of our experiment:

Is a use of shortcuts right after learning them is representative of a real context of use?

In a real context of use, we have to remember shortcuts that we learned in the past. For a better external validity, our design a protocol that includes two sessions on two consecutive days (learnability + memorability)

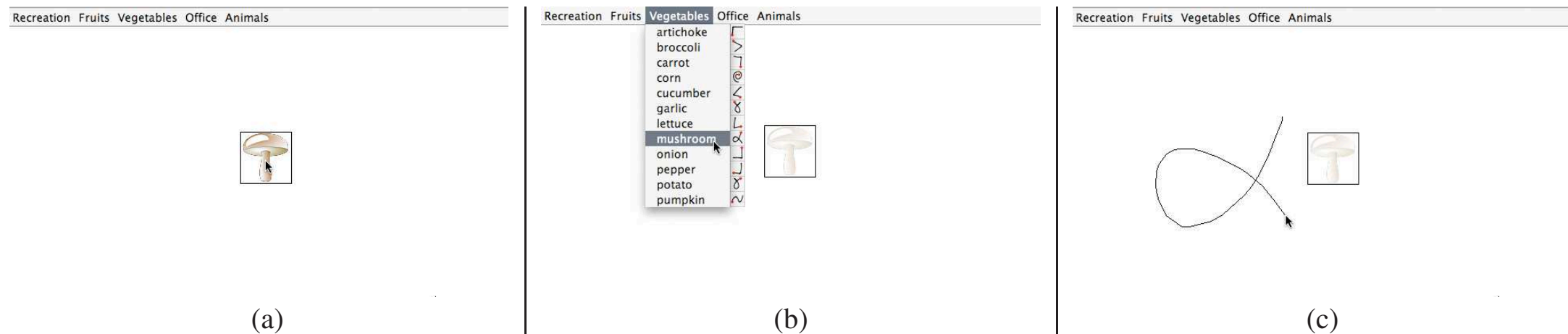
| | | |
|----------|---|-----------------|
| None |  | Warm up (day 1) |
| Keyboard |  | Test (day 1) |
| Stroke |  | Test (day 1) |
| Both |  | Re-test (day 2) |




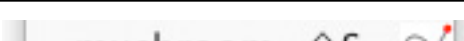
Operationalization

No standard defined - Example#1

Learning Keyboard vs Gesture shortcuts

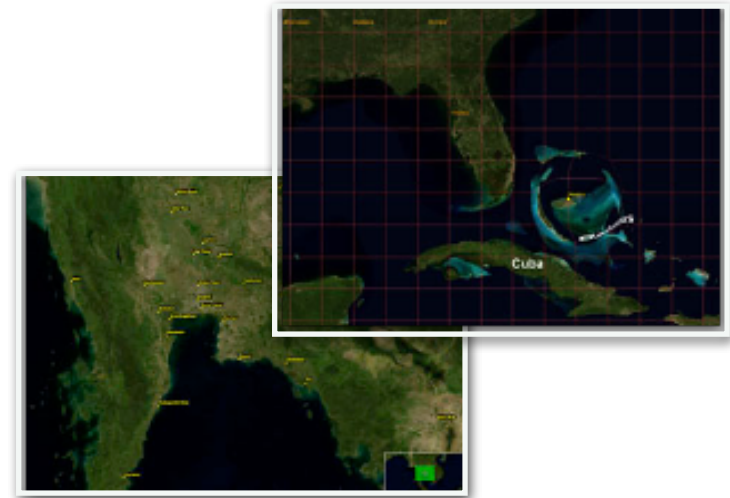
measure recall on two consecutive days
(learnability + memorability)



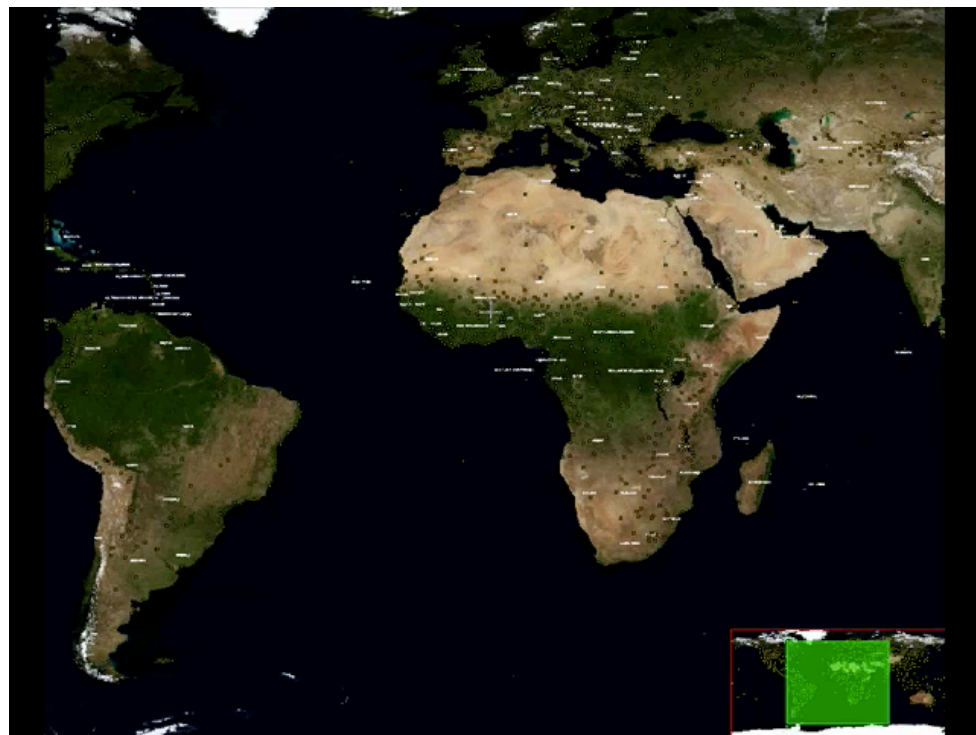
| | | |
|----------|---|-----------------|
| None |  | Warm up (day 1) |
| Keyboard |  | Test (day 1) |
| Stroke |  | Test (day 1) |
| Both |  | Re-test (day 2) |

Operationalization

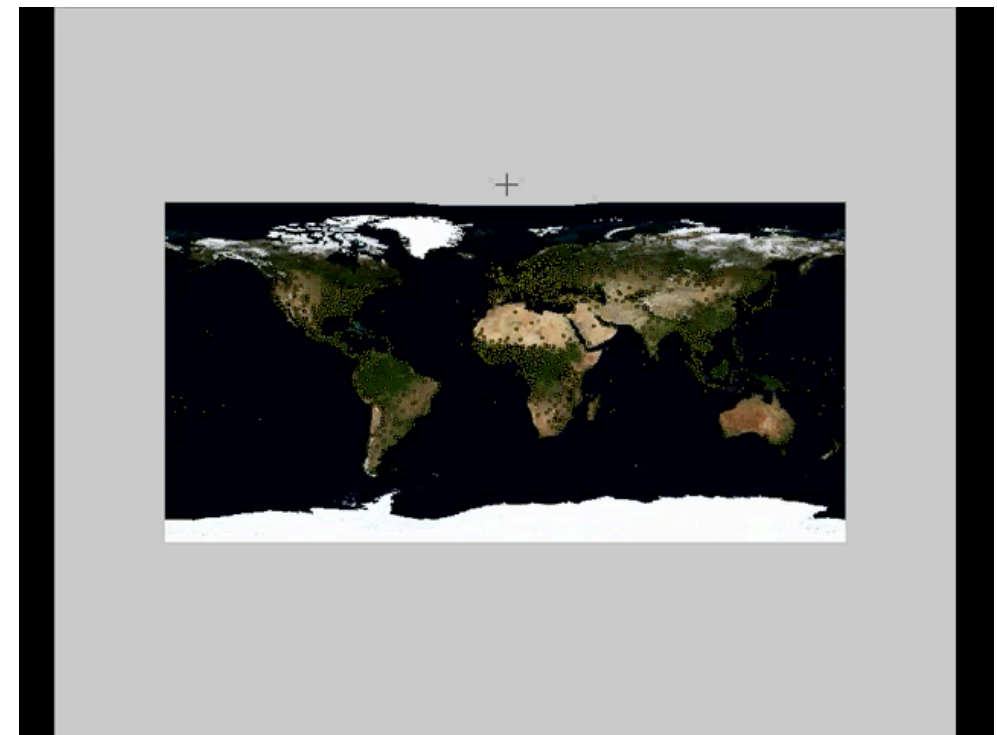
No standard defined - Example#2



H: The overview+detail navigation technique is more efficient than a magnifying lens to search an object in a zoomable space



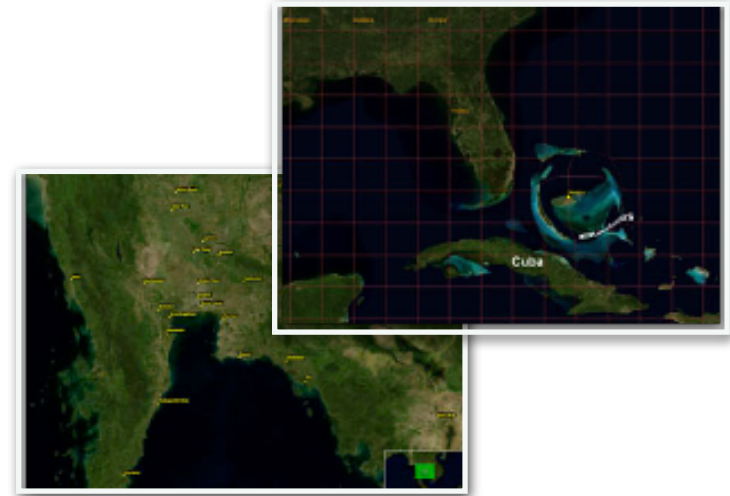
overview+detail



magnifying lens

Operationalization

No standard defined - Example#2



H: The overview+detail navigation technique is more efficient than a magnifying lens to search an object in a zoomable space

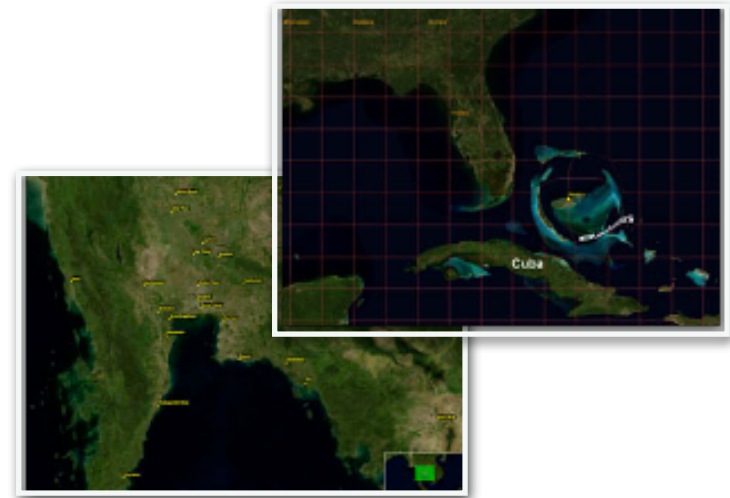
Factor: Navigation Technique {Overview, Lens}

Measure: Time to find a target object

Task: Ask the participant to search for a specific object that requires some zooming to be seen with the proper level of detail.

Operationalization

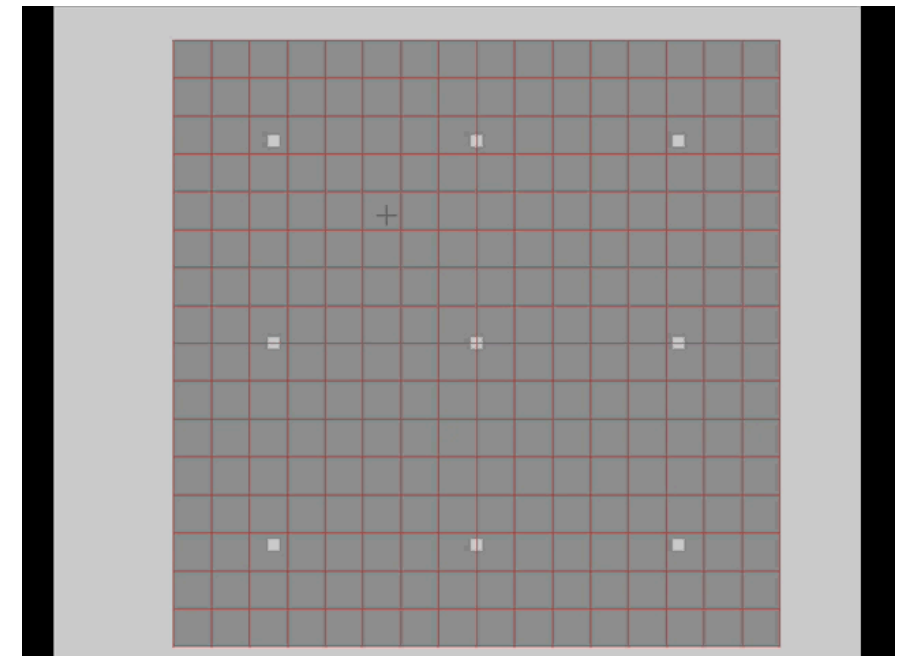
No standard defined - Example#2



Example of questions to assess the validity of our experiment:

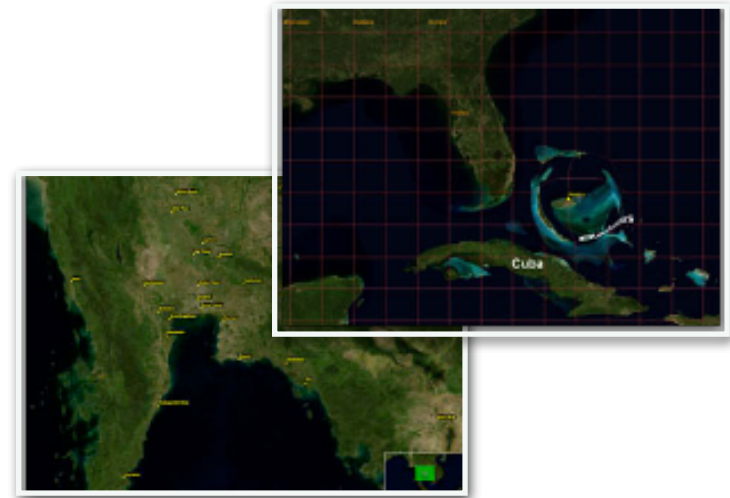
Should we use a real map with e.g., a target city to find?

No because of prior experience that we cannot easily control. For a better internal validity, let's rather use an abstract zoomable space. The target object to search is the one that has rounded corners.



Operationalization

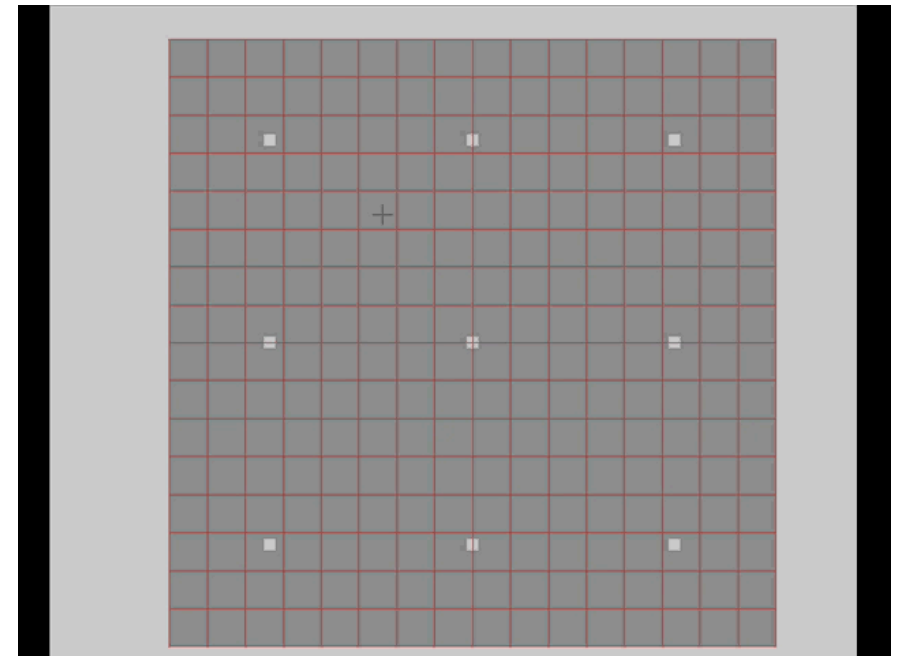
No standard defined - Example#2



Example of questions to assess the validity of our experiment:

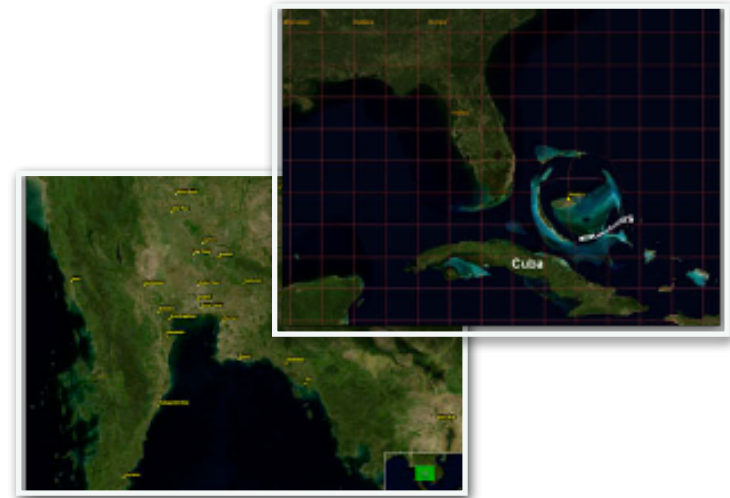
How can we make sure that participants will be forced to apply the same amount of zoom to see the target

There are individual differences in visual acuity that can threaten internal validity. We introduce an explicit action to unveil corners (press space bar). This action is enabled only at a zoom level where all participants can easily perceive rounded corners (need pilot studies).



Operationalization

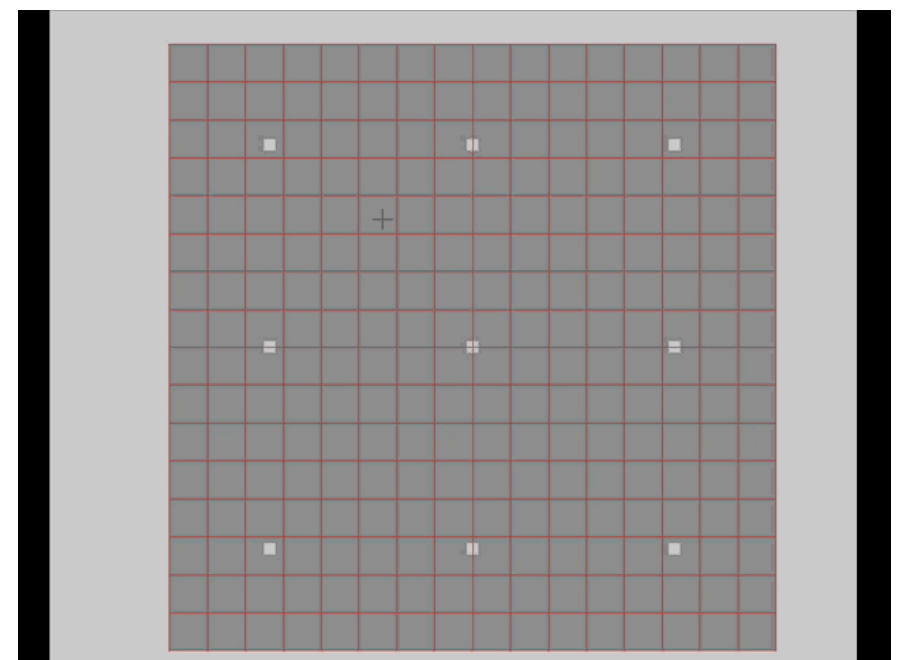
No standard defined - Example#2



Example of questions to assess the validity of our experiment:

How can we make sure that participants will be forced to navigate as much with one technique than with the other technique

There is a chance factor that can threaten internal validity if we randomly decide which object has rounded corners. We force a quantity of exploration a priori (at least n objects to inspect) with software control. This is easy to do thanks to the explicit action to inspect an object that we can rely on to count inspected objects.



Protocol design

Types of design

Choose a representative sample of the population you want to study.

How to assign and present the different experimental conditions?

If our experiment compares the mouse and trackpad conditions, which participants will test pointing with a mouse and/or pointing with a trackpad?

Choosing a type of design answers this question

Type of design: between- vs within

If the factor has n values, how do I present these n values?

between-subject design

Randomly select n groups of participants and assign a different factor value to each group

Assumptions

Other non-controlled variables are randomly distributed between the n groups

The only systematic difference between the n groups is the independent variable

within-subject design

Successively expose each participant to the n different values

Automatically controls most of the other variables

Allows experimenter to use a smaller number of participants

Best choice
if possible

Type of design: between- vs within

If the factor has n values, how do I present these n values?

Best choice
if possible

between-subject design

Participant 1

| Pointing device | Difficulty (ID) |
|-----------------|-----------------|
| Trackpad | 5 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 3 |
| Trackpad | 5 |
| Trackpad | 3 |
| Trackpad | 4 |
| Trackpad | 2 |
| Trackpad | 3 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 5 |

Participant 2

| Pointing device | Difficulty (ID) |
|-----------------|-----------------|
| Mouse | 2 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 4 |
| Mouse | 2 |
| Mouse | 4 |
| Mouse | 5 |
| Mouse | 3 |
| Mouse | 4 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 2 |

Pointing device is tested according to a between-subject design

within-subject design

Participant 1

| Pointing device | Difficulty (ID) |
|-----------------|-----------------|
| Trackpad | 5 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 3 |
| Trackpad | 5 |
| Trackpad | 3 |
| Trackpad | 4 |
| Trackpad | 2 |
| Trackpad | 3 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 5 |
| Mouse | 5 |
| Mouse | 2 |
| Mouse | 4 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 3 |
| Mouse | 4 |
| Mouse | 2 |
| Mouse | 3 |
| Mouse | 2 |
| Mouse | 4 |
| Mouse | 5 |

Participant 2

| Pointing device | Difficulty (ID) |
|-----------------|-----------------|
| Mouse | 2 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 4 |
| Mouse | 2 |
| Mouse | 4 |
| Mouse | 5 |
| Mouse | 3 |
| Mouse | 4 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 2 |
| Trackpad | 2 |
| Trackpad | 3 |
| Trackpad | 5 |
| Trackpad | 4 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 5 |
| Trackpad | 3 |
| Trackpad | 4 |
| Trackpad | 3 |
| Trackpad | 5 |
| Trackpad | 2 |

Pointing device is tested according to a within-subject design

Factorial design

Test several independent variables (factors) in the same experiment.

Mouse vs. Trackpad: device (mouse, trackpad), pointing difficulty (2, 3, 4, 5)

Each factor can be distributed according to a between or within subject design.

Experimental conditions

The number of experimental conditions depends on the number of factors and the different values these factors can take.

If there is **only one factor**, the number of experimental conditions is the number of possible values for this factor

```
2 Pointing device {Mouse, Trackpad}  
= 2 conditions
```

If there is **more than one factor**, there are as many experimental conditions as there are combinations of factor values

```
2 Pointing device {Mouse, Trackpad}  
x 4 Difficulty {2, 3, 4, 5}  
= 8 conditions
```

Each factor can be presented according to a between or within subject design (factorial design)

Type of design: between- vs within

If the factor has n values, how do I present these n values?

Best choice
if possible

between-subject design

Participant 1

| Pointing device | Difficulty (ID) |
|-----------------|-----------------|
| Trackpad | 5 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 3 |
| Trackpad | 5 |
| Trackpad | 3 |
| Trackpad | 4 |
| Trackpad | 2 |
| Trackpad | 3 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 5 |

Participant 2

| Pointing device | Difficulty (ID) |
|-----------------|-----------------|
| Mouse | 2 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 4 |
| Mouse | 2 |
| Mouse | 4 |
| Mouse | 5 |
| Mouse | 3 |
| Mouse | 4 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 2 |

*Pointing device is tested according to a between-subject design
Difficulty is tested according to a within-subject design*

within-subject design

Participant 1

| Pointing device | Difficulty (ID) |
|-----------------|-----------------|
| Trackpad | 5 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 3 |
| Trackpad | 5 |
| Trackpad | 3 |
| Trackpad | 4 |
| Trackpad | 2 |
| Trackpad | 3 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 5 |
| Mouse | 5 |
| Mouse | 2 |
| Mouse | 4 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 3 |
| Mouse | 4 |
| Mouse | 2 |
| Mouse | 3 |
| Mouse | 2 |
| Mouse | 4 |
| Mouse | 5 |

Participant 2

| Pointing device | Difficulty (ID) |
|-----------------|-----------------|
| Mouse | 2 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 4 |
| Mouse | 2 |
| Mouse | 4 |
| Mouse | 5 |
| Mouse | 3 |
| Mouse | 4 |
| Mouse | 3 |
| Mouse | 5 |
| Mouse | 2 |
| Trackpad | 2 |
| Trackpad | 3 |
| Trackpad | 5 |
| Trackpad | 4 |
| Trackpad | 2 |
| Trackpad | 4 |
| Trackpad | 5 |
| Trackpad | 3 |
| Trackpad | 4 |
| Trackpad | 3 |
| Trackpad | 5 |
| Trackpad | 2 |

*Pointing device is tested according to a within-subject design
Difficulty is tested according to a within-subject design*

Controlling variation

Replication and blocking are two mechanisms to reduce observed variation that is not due to difference between conditions

Replication: A participant does several times the experimental task in the same condition.

e.g. if the participant got distracted in a particular condition

Blocking: Arranging the tasks into blocks of tasks that are similar to one another.

e.g. eliminating the time due to successive changes between two pointing devices

Controlling variation

Participant 1

| | Pointing device | Pointing Difficulty (ID) |
|-------------------|-----------------|--------------------------|
| Trackpad block | Trackpad | 5 |
| | Trackpad | 2 |
| | Trackpad | 4 |
| | Trackpad | 3 |
| | Trackpad | 5 |
| | Trackpad | 3 |
| | Trackpad | 4 |
| | Trackpad | 2 |
| | Trackpad | 3 |
| | Trackpad | 2 |
| | Trackpad | 4 |
| | Trackpad | 5 |
| Mouse block | Mouse | 5 |
| | Mouse | 2 |
| | Mouse | 4 |
| | Mouse | 3 |
| | Mouse | 5 |
| | Mouse | 3 |
| | Mouse | 4 |
| | Mouse | 2 |
| | Mouse | 3 |
| | Mouse | 2 |
| | Mouse | 4 |
| | Mouse | 5 |

Participant 2

| | Pointing device | Pointing Difficulty (ID) |
|-------------------|-----------------|--------------------------|
| Mouse block | Mouse | 2 |
| | Mouse | 3 |
| | Mouse | 5 |
| | Mouse | 4 |
| | Mouse | 2 |
| | Mouse | 4 |
| | Mouse | 5 |
| | Mouse | 3 |
| | Mouse | 4 |
| | Mouse | 3 |
| | Mouse | 5 |
| | Mouse | 2 |
| Trackpad block | Trackpad | 2 |
| | Trackpad | 3 |
| | Trackpad | 5 |
| | Trackpad | 4 |
| | Trackpad | 2 |
| | Trackpad | 4 |
| | Trackpad | 5 |
| | Trackpad | 3 |
| | Trackpad | 4 |
| | Trackpad | 3 |
| | Trackpad | 5 |
| | Trackpad | 2 |

*Trials are blocked per Pointing Device condition
Each condition is replicated three times (e.g. in yellow, 3 x (Trackpad x ID))*

Controlling order effect

Order effects happen when an independent variable (factor) is presented according to a within-subject design

if the mouse is always presented after the trackpad and observed time is shorter with the mouse, is pointing performance better because of the input device or because the user has become more familiar (thus efficient) with the task?

Randomization means presenting the different conditions in a “random” order across the experiment

Controlling order effect

Participant 1

| | Pointing device | Pointing Difficulty (ID) |
|----------------|-----------------|--------------------------|
| Trackpad block | Trackpad | 5 |
| | Trackpad | 2 |
| | Trackpad | 4 |
| | Trackpad | 3 |
| | Trackpad | 5 |
| | Trackpad | 3 |
| | Trackpad | 4 |
| | Trackpad | 2 |
| | Trackpad | 3 |
| | Trackpad | 2 |
| | Trackpad | 4 |
| | Trackpad | 5 |
| Mouse block | Mouse | 5 |
| | Mouse | 2 |
| | Mouse | 4 |
| | Mouse | 3 |
| | Mouse | 5 |
| | Mouse | 3 |
| | Mouse | 4 |
| | Mouse | 2 |
| | Mouse | 3 |
| | Mouse | 2 |
| | Mouse | 4 |
| | Mouse | 5 |

Participant 2

| | Pointing device | Pointing Difficulty (ID) |
|----------------|-----------------|--------------------------|
| Mouse block | Mouse | 2 |
| | Mouse | 3 |
| | Mouse | 5 |
| | Mouse | 4 |
| | Mouse | 2 |
| | Mouse | 4 |
| | Mouse | 5 |
| | Mouse | 3 |
| | Mouse | 4 |
| | Mouse | 3 |
| | Mouse | 5 |
| | Mouse | 2 |
| Trackpad block | Trackpad | 2 |
| | Trackpad | 3 |
| | Trackpad | 5 |
| | Trackpad | 4 |
| | Trackpad | 2 |
| | Trackpad | 4 |
| | Trackpad | 5 |
| | Trackpad | 3 |
| | Trackpad | 4 |
| | Trackpad | 3 |
| | Trackpad | 5 |
| | Trackpad | 2 |

Participant 1 starts with the Trackpad condition
Participant 2 starts with the Mouse condition
=> Presentation order for Pointing Device is randomized

Randomization

Randomization is not haphazard

An experiment is randomized if the method for assigning levels of independent variables involves a deterministic probabilistic scheme.

Example of bad randomization

assign mouse or trackpad depending on if start time in seconds is a odd or even (pb: can result in much more observations in one or the other condition)

Counterbalancing

Counterbalancing is a scheme to randomize a within-subject experiment design

Consider a factor that has n levels and a sample of X participants, we have three possible strategies:

Complete: Compute the $n!$ possible orders and assign $X / n!$ participants to each order

-- requires a multiple of $n!$ participants

Latin Square: Compute n possible orders using a Latin Square and assign X / n participants to each order

-- requires a multiple of n participants

Random: Compute m (potentially $< n$) orders using a randomized algorithm and assign X / m participants to each order

-- requires a multiple of m participants

⇐ use with caution

Randomization

Latin square definition

A Latin square is an $n \times n$ array filled with n different symbols, each occurring exactly once in each row and exactly once in each column

Example: $n=3$ levels ($\{A, B, C\}$)

| | | |
|---|---|---|
| A | B | C |
| C | A | B |
| B | C | A |

Ensures that the two orders between each possible pair are represented

$A \rightarrow B$, $A \rightarrow C$, $B \rightarrow A$, $C \rightarrow A$

(but elements in each pair might not be consecutive)

Randomization

A concrete example

An experiment entails comparing four input devices regarding their pointing performance

Factor: {Mouse, TrackPad, Pen, Finger}

Measure: Completion time

Task: Point a target

Using a **Complete** strategy to counterbalance the presentation order of experimental conditions would require 24 (4!) participants. We can't afford it, we have access to 15 participants max. We rather use a **Latin Square** to compute 4 representative orders, assign 3 participants to each order and recruit only 12 participants [P1, ..., P12].

| | | | | | |
|---------|----------|----------|----------|----------|---------------|
| Order 1 | Mouse | TrackPad | Pen | Finger | P1, P2, P3 |
| Order 2 | TrackPad | Finger | Mouse | Pen | P4, P5, P6 |
| Order 3 | Finger | Pen | TrackPad | Mouse | P7, P8, P9 |
| Order 4 | Pen | Mouse | Finger | TrackPad | P10, P11, P12 |

Replicability

Any experiment should be **replicable** by others

Always report:

- the experiment's goal
- the hardware/software characteristics of the experimental environment (apparatus)
- a description of the participants' characteristics that may impact the observed measures (gender, mean and variance in age, prior experience...)
- a complete description of the experimental task
- a complete description of the experiment procedure (experiment design and the main steps each participant went through)

EXPERIMENT 1: FOCUS TARGETING PERFORMANCE

We conducted an experiment to compare the performance and limits of the three existing and two new lenses described in the previous section. Participants were asked to perform a

...

Apparatus

We used a Dell Precision 380 equipped with a 3 GHz Pentium D processor, an NVidia Quadro FX4500 graphics card, a 1600 x 1200 LCD monitor (21") and a Dell optical mouse. The program was written in Java 1.6 using the open source ZVTM toolkit [23] which offers a wide range of distortion lenses and could easily be extended to support translucence- and time-based transitions. The application was limited to a 1400 x 1200 window with a black padding of 100 pixels in order to accommodate instruction messages and simulate screen real-estate that would usually be taken by control and information widgets.

Participants

Ten unpaid adult volunteers (7 male, 3 female), from 23 to 40 year-old (average 26.4, median 25), all with normal or corrected to normal vision, served in the experiment.

Task and Procedure

Our focus targeting task consisted in acquiring a target in the flat-top of the lens as quick as possible. In our experimental setting, the lens was centered on the mouse cursor. The task ended when the participant clicked the left mouse but-

...

Running the experiment

Ethics

Testing can be a distressing experience

- pressure to perform, errors inevitable

- feelings of inadequacy

- competition with other participants

Golden rules

- participants should always be treated with respect

- always explain you are testing the system, not the user

- explain how comments and criticisms are good

Running the experiment

Control for bias, avoid Hawthorne effect

Hawthorne effect: changes in participants' behavior during the course of a study may be "related only to the special social situation and social treatment they received."

unbiased instructions (write them down before)

double-anonymous if possible (the experimenter and the participant do not know which group it is). This is frequent in medicine but, in HCI, it is rather rare. In all cases, the participant does not know what the hypothesis is.

Keep the same conditions (software, environment) from one participant to another

Before the test

Get approval from your IRB (Institutional Review Board)

Many institutions have an ethics committee for any experiment that involves human participants

They usually review the experiment's purpose, the protocol, the recruitment process, the consent form to give to participants, data policies, etc.

Don't waste participants' time

debug and set up the experiment environment

make participants feel comfortable

acknowledge that the software may have problems

let participants know they can stop at any time

Consent form

Only use volunteers

Participants should **sign a consent form** that explains any monitoring that is being used

Maintain privacy

tell participant that individual test results are confidential

explain any monitoring that is being used

INFORMATION NOTICE AND INFORMED CONSENT EXPERIMENT 1

Project title: Virtual duplication of collaborator's body and display in a multi-display environment

Researcher in charge of the project:

Caroline Appert, caroline.appert@universite-paris-saclay.fr, 0169153460, Bâtiment 660, Université Paris-Saclay

Where the experiment takes place: Bâtiment 660, Université Paris-Saclay

Goal of the research project: Development of interaction techniques to facilitate communication in a multi-display environment.

What we expect from you:

We are interested in the collaboration of touch screen users when the layout or orientation of these screens makes communication difficult. This technique is based on the use of "virtual duplicates", which are virtual objects visible in Augmented Reality representing the body and screen of a collaborator. During this experience, you will be equipped with a "video pass-through" Virtual Reality headset that allows you to see the physical environment around you through the headset's screens, while displaying virtual elements "in the air". The Virtual Reality headset used will be a ~~Varjo~~ XR-3 (<https://varjo.com/products/xr-3/>). We'll ask you to carry out a series of tasks in which you'll use a touch screen and be asked to use virtual duplicates to collaborate with the experimenter. You will test four conditions in which the duplicate of the experimenter's body will be represented either abstractly, by an avatar reduced to a view pyramid and a pointer, or concretely, by a volumetric avatar representing the head, torso, and hands, in a position that is either physically plausible or unreal (a position that is physically impossible or does not respect the social conventions of proximity). There's no such thing as the "best configuration"; our aim is to understand the advantages and disadvantages of each. During and at the end of the experiment, we'll ask you to rate each configuration in terms of perceived ease and efficiency in carrying out the tasks.

Your rights to withdraw from the research at any time:

Your contribution to this research is voluntary. You can stop your participation without any justification at any time. You simply tell the experimenter that you want to stop. Withdrawing from the study can in no way influence future relations with the researchers in charge of this study, the LISN laboratory, or University Paris-Saclay.

Your rights to confidentiality and privacy:

The data obtained will be treated with the utmost confidentiality. We will mask your identity with a random number and no other information will reveal your identity. The consent form is stored in paper format in a locked drawer. Post-experiment questionnaires will be analyzed within a month of the experiment and then destroyed. The software records the execution time and the number of errors during the tasks. These data are not associated with any personal information (completely anonymized) and will be made public for the scientific community. We may use the comments you make during the study to explain our results in a scientific paper. The comments will be completely anonymous. If you do not want a comment to be used, just tell the experimenter.

Benefits of the study:

Propose new ways of interacting in Augmented Reality for multi-display collaboration.

Possible risks of the study:

There is no risk a priori, but if you feel any discomfort (such as fatigue, nausea or dizziness), notify the operator immediately. The experiment will stop. In any case, you can take a break or stop completely at any time, without any justification.

Dissemination:

This research will be disseminated at scientific conferences and will be published in conference proceedings and academic journal articles.

Your right to ask questions:

There is no risk a priori, but if you feel any discomfort (such as fatigue, nausea, or dizziness), notify the operator immediately. The experiment will stop. In any case, you can take a break or stop completely at any time, without any justification. If you are pregnant or suffering from epilepsy, motion sickness, migraines, or imbalance problems, you must not take part in this experiment.

Consent to participate:

By signing the consent form, you certify that you have read and understood the above information, that the researcher has answered your questions satisfactorily, and that the researcher has advised you that you are free to withdraw your consent or withdraw from this research at any time without prejudice.

To be completed by the participant:

I have read and understand the above information and willingly agree to participate in this research.

Date, Last Name, First Name, Signature:

To be completed by the experimenter:

Date, Last Name, First Name, Signature:

After the test

Make participants feel comfortable

state that the participant has helped you find areas of improvement

inform the participant about what hypotheses you are testing

answer particular questions about the experiment that could have biased the results before

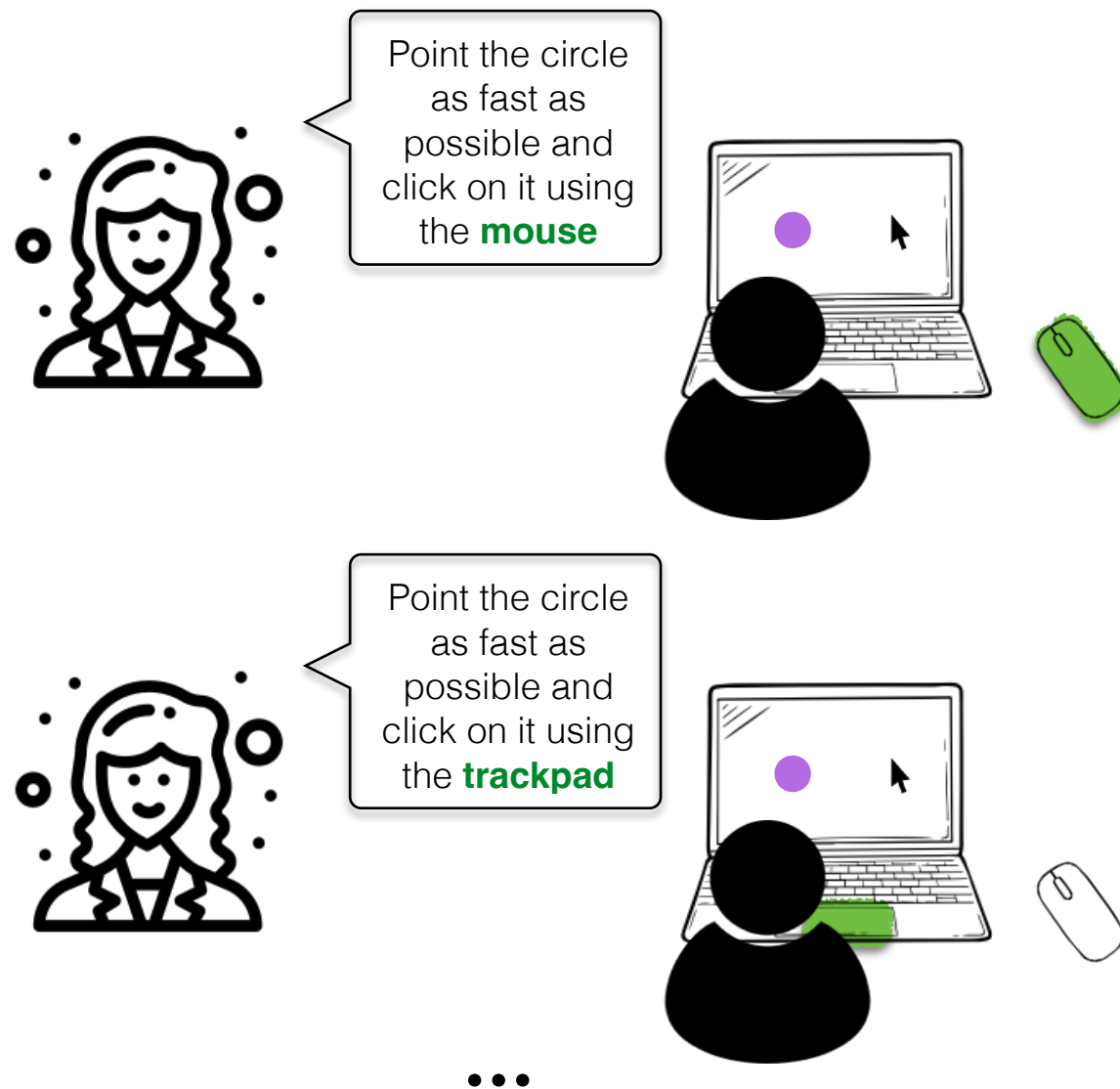
Maintain privacy

never report results in a way that individual participants can be identified

only show videotapes outside the research group **with participants' permission**

Laboratory experiment - overview

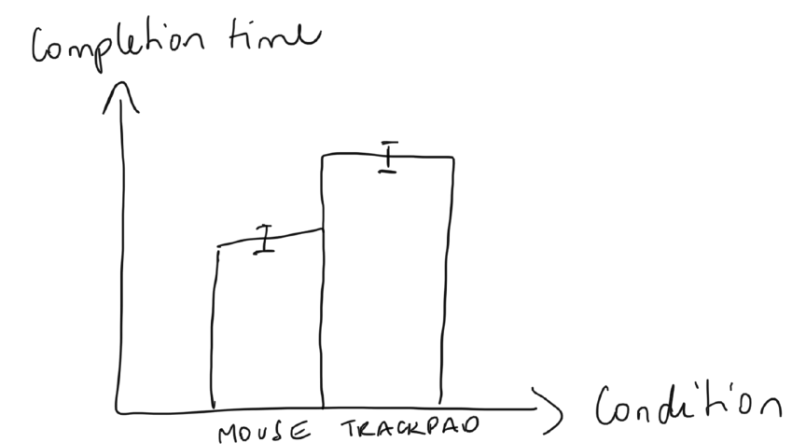
Hypothesis: Users point faster with a mouse than with a trackpad



1. The operator asks participants to complete tasks under **specific conditions**

| Participant | Condition | Completion time |
|-------------|-----------|-----------------|
| P1 | mouse | 403 |
| P1 | trackpad | 527 |
| ⋮ | ⋮ | ⋮ |
| P2 | mouse | 522 |
| P2 | trackpad | 608 |
| ⋮ | ⋮ | ⋮ |

2. Participants' performance is recorded in **log files**



3. Log files are analyzed with statistical procedures to test the research hypothesis

Recording measures (data logging)

Save one log file per participant in tabular format with one line per run task (trial-level). Each line describes a trial: general info, factor values, and measure values.

```
Participant,Practice,Block,Trial,Device,Difficulty,PointingTime
0,true,0,0,Trackpad,Easy,1632
0,true,0,1,Trackpad,Medium,1552
0,true,0,2,Trackpad,Hard,2030
0,false,1,0,Trackpad,Hard,1582
0,false,1,1,Trackpad,Medium,1639
...
11,false,3,19,Mouse,Easy,1582
11,false,3,20,Mouse,Hard,1639
```

General info

Factors

Measures

Save a log file that is easy to analyze by humans and machines

detail each acronym/short name you may use in your log files (*e.g.*, TP means TrackPad)

log the run date directly in the file with a dedicated column or in the file name (*e.g.*, log_P1_2023_01_15_14h52.csv)

Recording measures (data logging)

Collect computed measures and raw data

to fix potential undetected bugs or allow you to do analyses that you did not plan in advance

Example: In a pointing experiment, you have a measure `hit={yes, no}` that you *compute* based on the cursor and target positions. Collecting *raw data* (`cursor_x`, `cursor_y`, `target_x`, `target_y`) in addition can allow you to get an estimation of the distance error when `hit=no` even if you had not planned it.

If relevant, complement the trial-level log file with an event-level log file

Example: In an experiment testing the accuracy of a gesture recognizer, collect the recognized gesture in your main trial-level log file. Collecting all (x,y,t) in a separate event-level file allows you to replay participants' input to test alternative gesture recognizers.