

human computer interaction

Lecture 7: Evaluating With Users

Dmitrijs Dmitrenko





Good design vs Bad design



bad

design: Norman door in India



bad

design: labelling of floors in a lift



Thanks, Saranya, for sharing this example!

bad
design: an
extreme case 😊



Thanks, Moldir, for sharing
this example!



share your observations...

...on X or per email!

Use hashtags **#HCISussex #GoodDesign #BadDesign**

Tag me @DoubleDmi

Your examples will be featured in lectures!

「Now back to today's
lecture...」

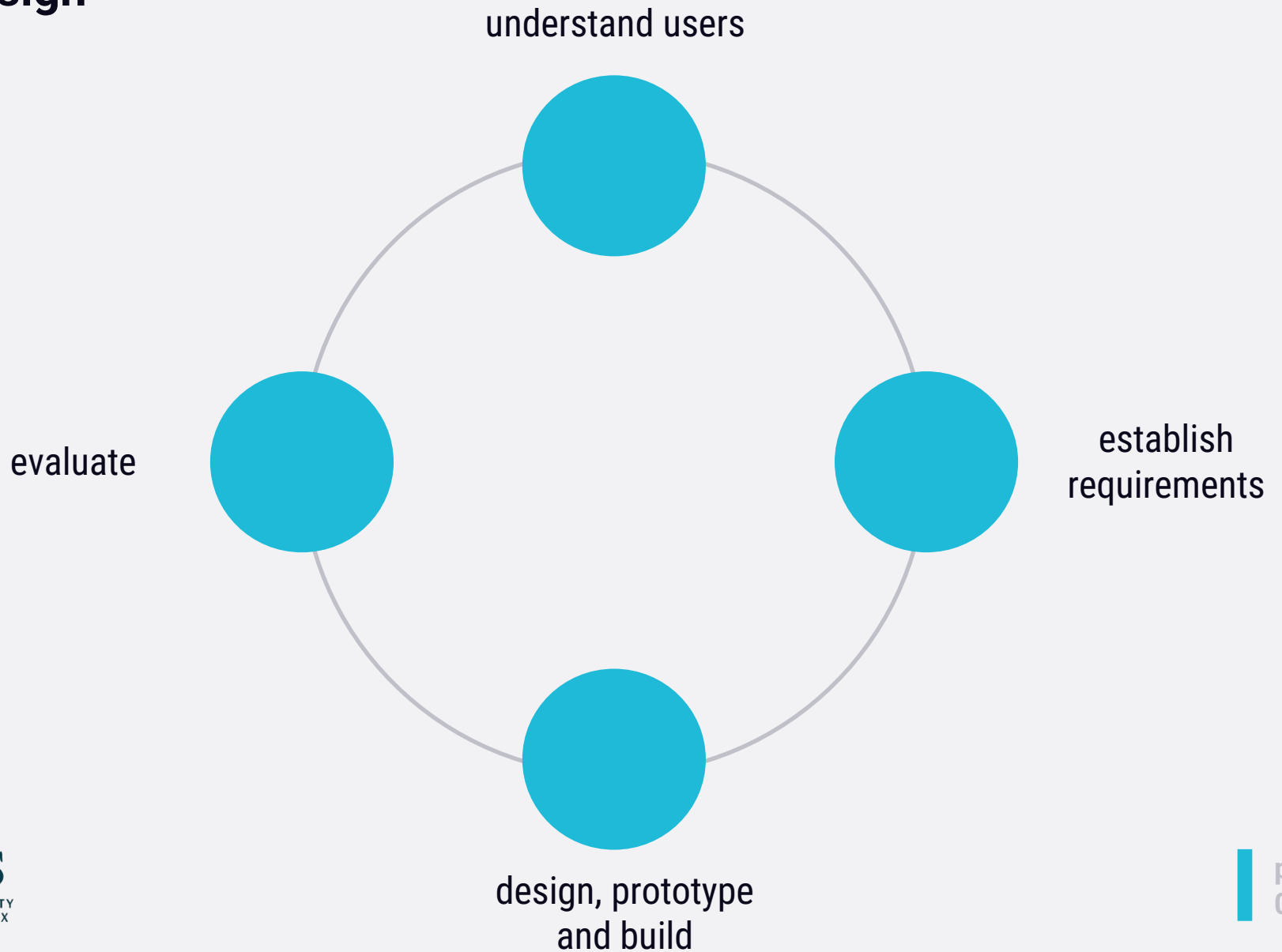
Devs watching QA test the product



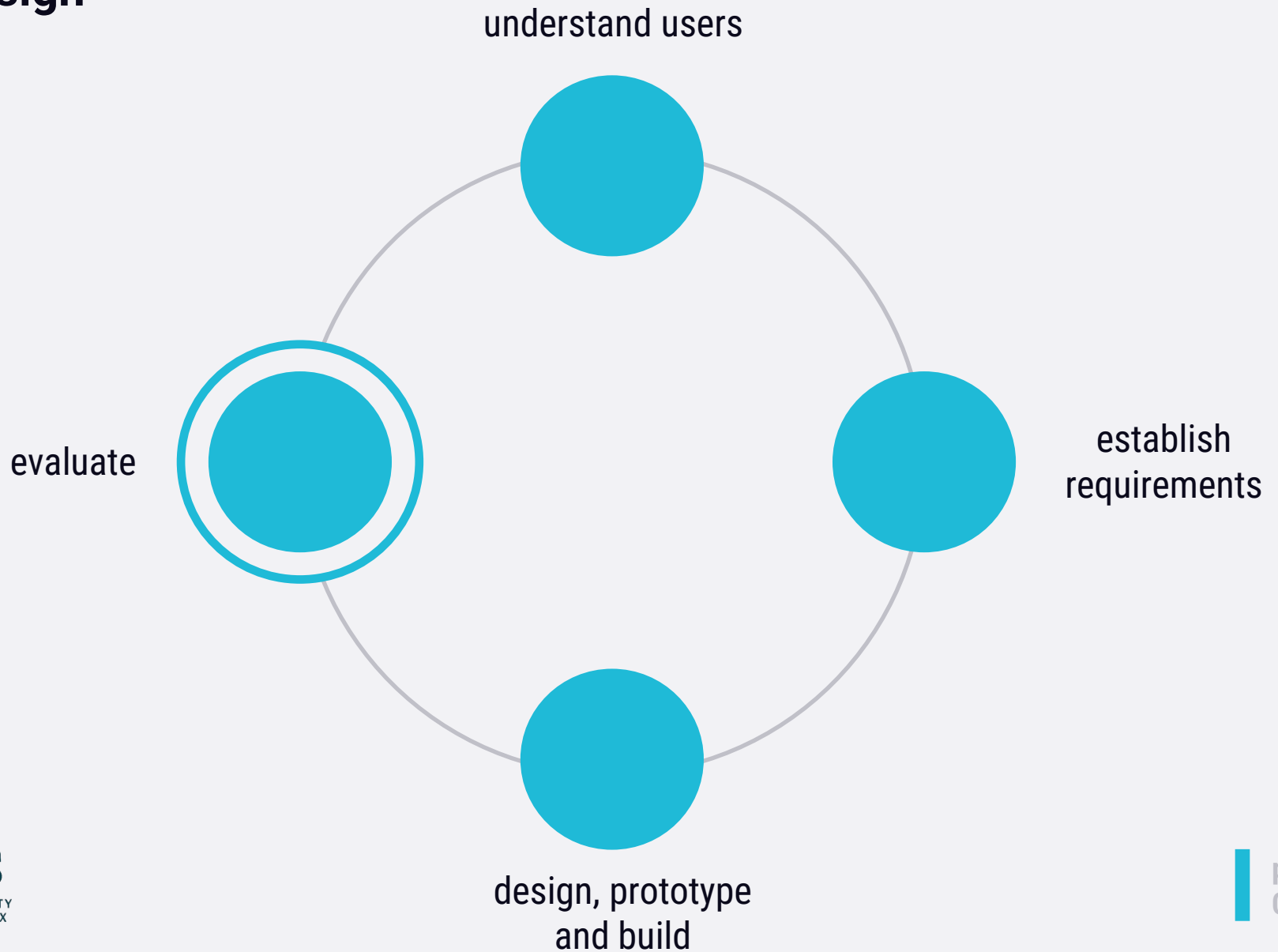


last week:
prototyping designs

user-centred design



user-centred design



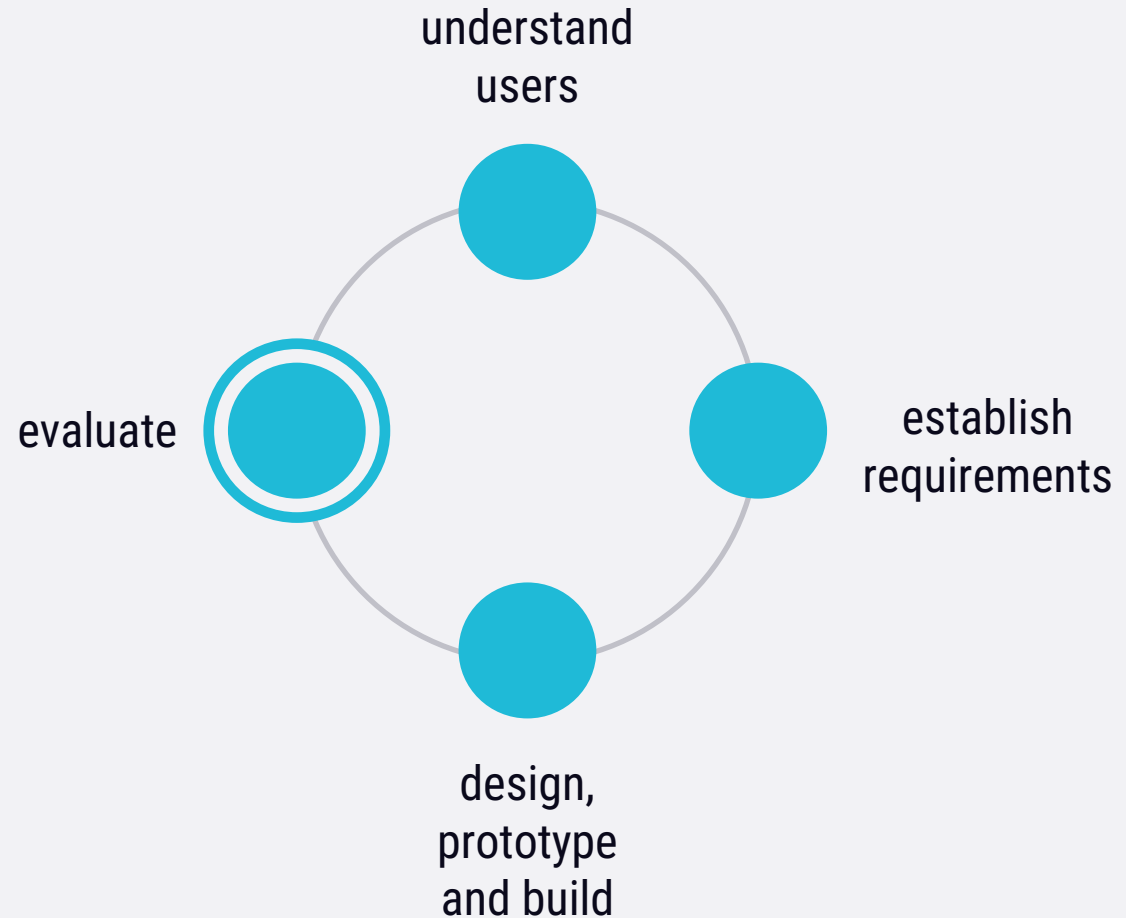
evaluating with users

in controlled
environments

- 🧑‍🔬 usability testing
- 🧑‍🔬 experiments

‘in the wild’

- 🧑‍🔬 field studies



evaluation in controlled environments





Usability Testing



usability testing

- ⌚ takes place in a **controlled setting**
- ⌚ focus on how well users perform tasks with the product
- ⌚ involves **testing** and recording **typical users' performance** on typical tasks
- ⌚ **users are observed** and timed
- ⌚ testers **record time taken** to complete task and **number/type of errors**
- ⌚ **comparison of products or prototypes is common**

usability testing: 'lab in a box'



usability testing: portable lab



Source: <https://twitter.com/ClemensScharti/media>

usability testing: data collection

🕒 data gathered through video & interaction logging (e.g. key presses) to record:

🕒 **time to complete a task**

🕒 time to complete a task after a specified time away from the product

🕒 **number** and type **of errors** per task

🕒 number of errors per unit of time

🕒 **number of navigations to online help** or manuals

🕒 number of users making a particular error

🕒 number of users **completing task successfully**

(mainly quantitative data)

🗨️ **interview and questionnaires** can add to this data and help to explain errors or frustrations

(mainly qualitative data)

usability testing: monday.com

What data would you collect?



usability testing: monday.com

🔄 possible quantitative data:

- 🔄 time to setup a profile or a team
- 🔄 time to create a task and assign it to a colleague
- 🔄 time to update the status of a task
- 🔄 time to define the timeline
- 🔄 time to check the overview on the Gantt chart
- 🔄 number and type of errors per task
- 🔄 number of errors per unit of time
- 🔄 number of navigations to online help or manuals
- 🔄 number of users making a particular error
- 🔄 number of users completing task successfully

🔄 possible interview/questionnaire questions?

- 🔄 What's the hardest part about using this product?
- 🔄 Was there anything surprising or unexpected about this product?
- 🔄 Was there anything missing from this product that you expected?

usability testing: conditions

- ④ emphasis on:
 - ④ selecting **representative users**
 - ④ developing **representative tasks**
- ④ informed **consent form** explains procedures and deals with **ethical issues**
- ④ the **test conditions** should be the **same for every participant**
- ④ tasks usually last no more than 30 minutes

usability testing: participants

- ⊕ how many participants?
- ⊕ practical issue depending on:
 - ⊕ schedule for testing
 - ⊕ availability of participants
 - ⊕ cost of running tests
- ⊕ 5-10 users typically selected
- ⊕ however, some experts argue that testing should continue until no new insights are gained

usability testing: pros & cons

pros

uninterrupted – can assess performance, identify errors and help **explain why users did what they did**



can use in conjunction with **satisfaction questionnaires and interviews** to elicit user opinions



cons



controlled settings are **artificial** and lack context



requires **skill to determine typical users** and typical tasks



typically requires (at least) 4 members of the design team



time to set up tests, recruit participants, and run tests



need access to **resources/equipment**



Experiments



experiments

- ④ formulate and **test hypothesis** (typing is faster than swiping)
- ④ predict the **relationship between** two or more **variables** (typing time, keyboard type)
- ④ **independent variable** is manipulated by the researcher (keyboard type, language)
- ④ **dependent variable** influenced by the independent variable (typing time)
- ④ typical experimental designs have one or two independent variables
- ④ sample size may be defined using [Power Analysis](#)
- ④ validated statistically & replicable

experiments

- ⊗ imagine testing which colour of a “Join Our Waiting List” button results in more users joining the list.

Join Our Waiting List

Join Our Waiting List

- ⊗ the colour of the button is your **independent variable** with two values (i.e. “blue” and “green”)
- ⊗ the number of people joining the list is your **dependent variable**
- ⊗ you are studying the **effect** of the button’s colour on the number of people joining the list

setting up an experiment: example

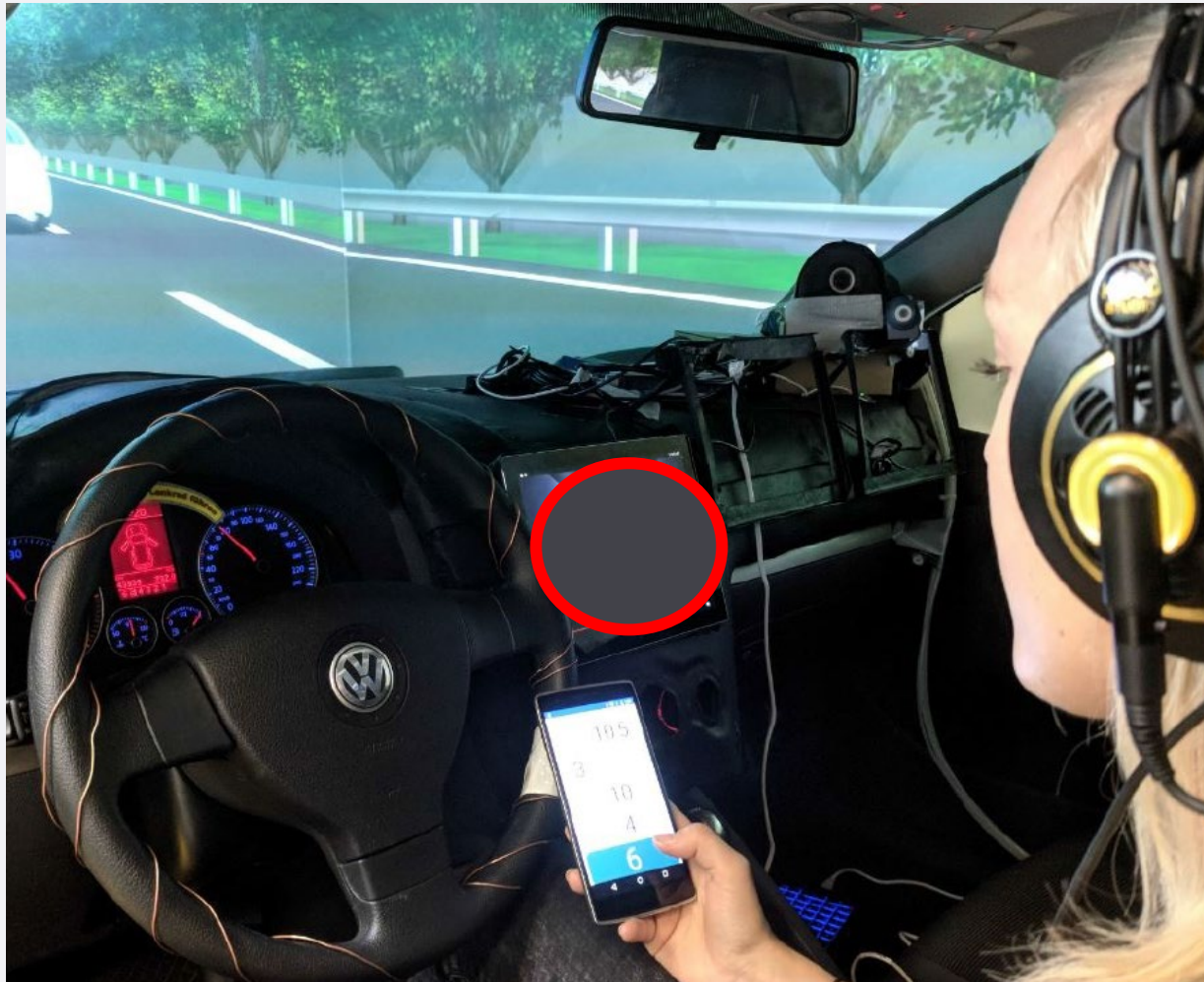


Source: <https://dl.acm.org/citation.cfm?id=3302332>

setting up an experiment: defining an independent variable



setting up an experiment: defining an independent variable



Source: <https://dl.acm.org/citation.cfm?id=3302332>

setting up an experiment: defining an independent variable



setting up an experiment: defining an independent variable

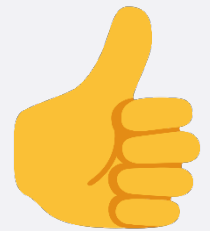


setting up an experiment: defining dependent variables



Source: <https://dl.acm.org/citation.cfm?id=3302332>

setting up an experiment: defining dependent variables



setting up an experiment: defining dependent variables



setting up an experiment: defining dependent variables

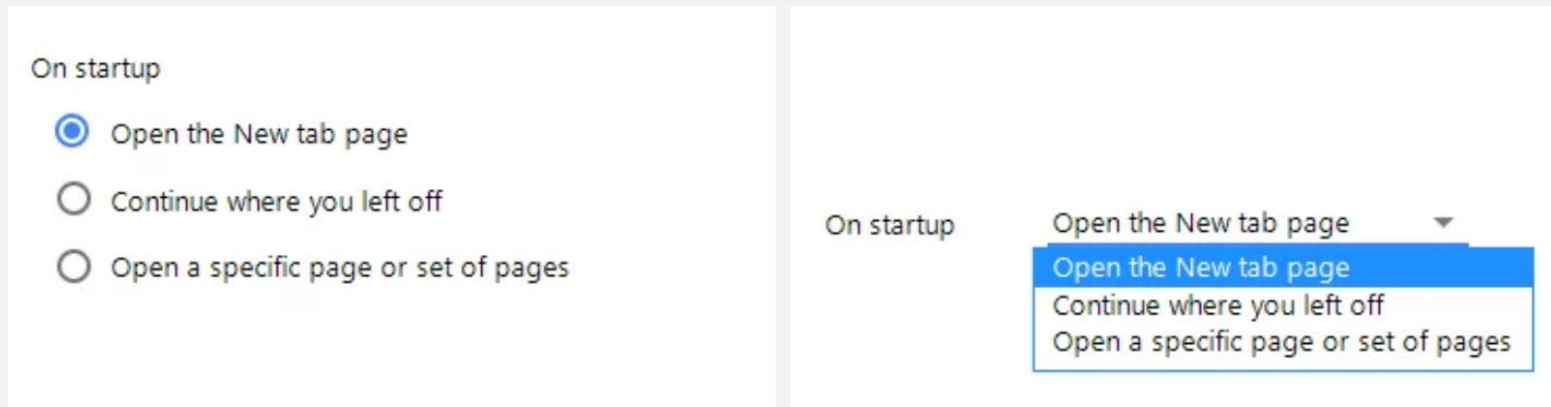


I trust it!



setting up an experiment: defining variables

- Imagine you are evaluating an application with two competing interfaces:



- Your task is to find out which one is easier to use.
- What would be your **dependent and independent variables**?

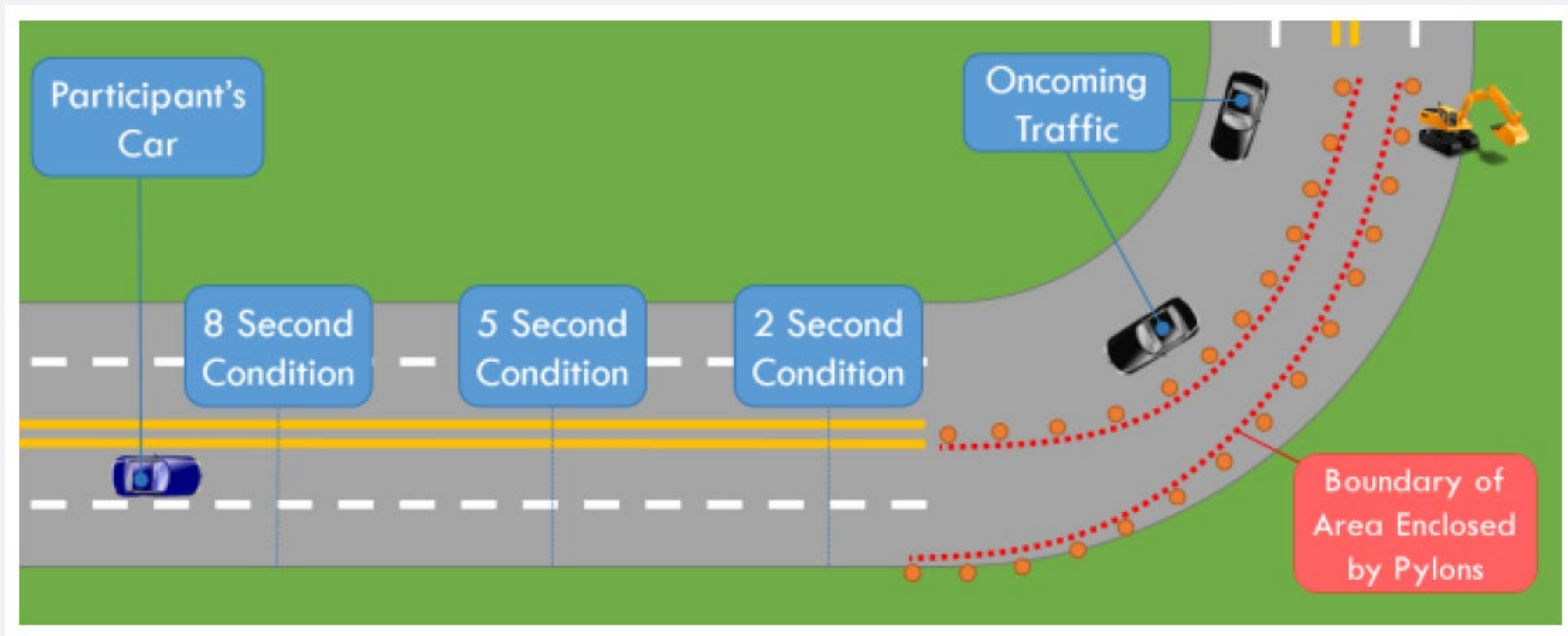
experimental designs

- ⊕ **different (between) participants:** single group of participants is allocated randomly to the experimental conditions
- ⊕ **same (within) participants:** all participants take part in both conditions

What **experimental design** should be used in this example?

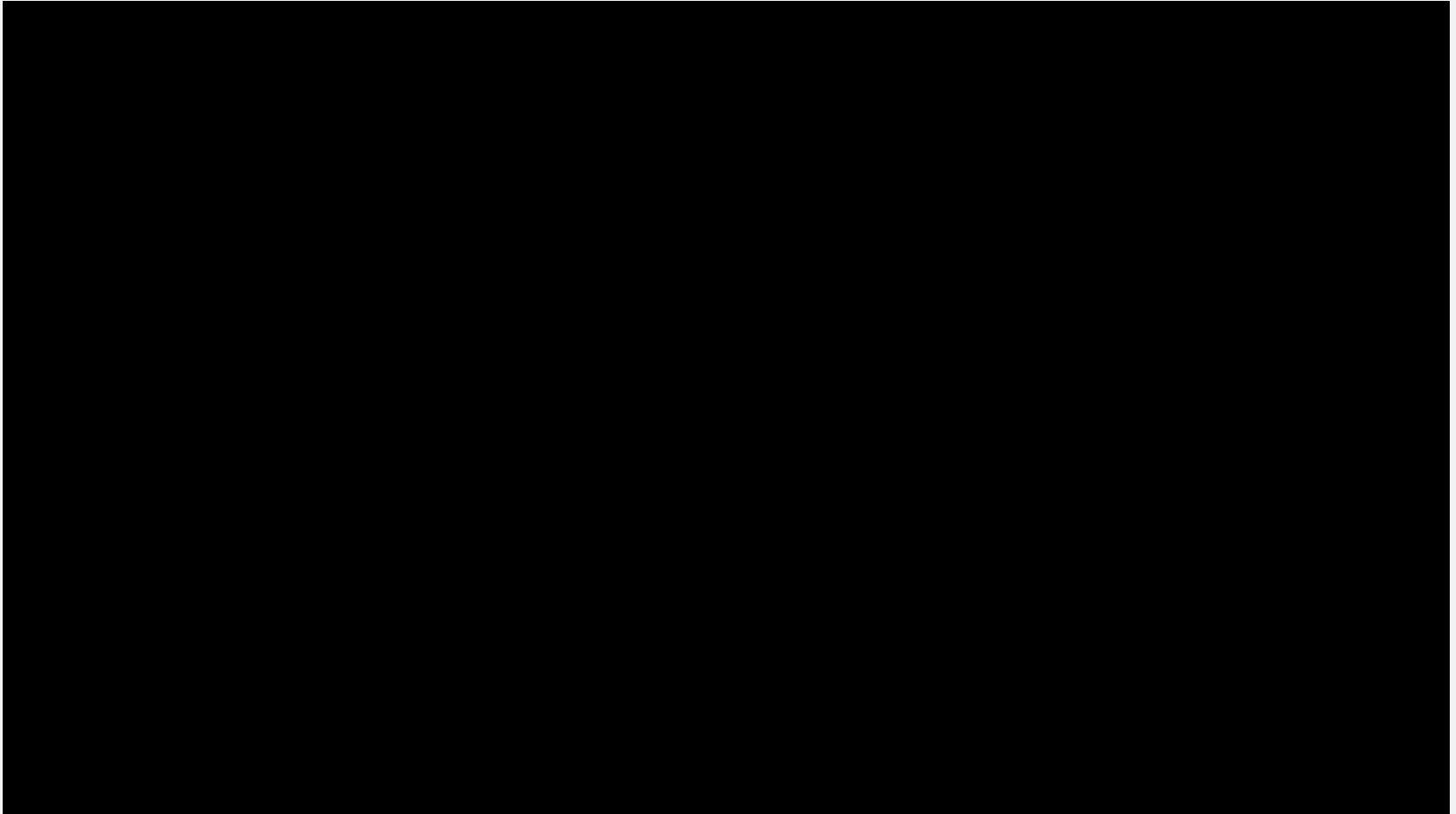
“We tested three transition time conditions, with an unstructured transition of control occurring 2s, 5s, or 8s before entering a curve.”

“Few drivers in the 2 second condition were able to safely negotiate the road hazard situation, while the majority of drivers in the 5 or 8 second conditions were able to navigate the hazard situation safely.”



Source: ([Mok, 2017](#))

example different participants



What **experimental design** should be used in this example?

“We ran a study to explore the effect of three notifications (conveyed either as visual only or combined visual-olfactory stimuli) on the number of driving-relevant mistakes and the perceived distraction, helpfulness, liking, and comfort. The scents of lavender, peppermint, and lemon were used accordingly. The order of the two conditions was randomised.”



experimental designs (overview)

design	pros	cons
different (between)	no order effects	many participants & individual differences
same (within)	few individuals, no individual differences	counterbalancing needed because of ordering effects

experiments vs usability testing

- ⊗ **experiments test hypotheses** to discover new knowledge by investigating the **relationship between two or more things** – i.e. variables
- ⊗ **usability testing is applied experimentation**
- ⊗ developers **check that** the system is **usable** by the intended user population for their tasks
- ⊗ developers often make predictions and hypotheses about how users will interact with the system(s)

experiments vs usability testing: using standardised questionnaires

- ⊗ **Semi-structured post-study interviews** are important for both usability tests & experiments
- ⊗ Consider using **standardised questionnaires**, e.g.:
 - ⊗ Technology Acceptance Model (TAM)
 - ⊗ NASA Task Load Index (TLX)
 - ⊗ System Usability Scale (SUS)
 - ⊗ Player Experience Inventory (PXI)
 - ⊗ User Experience Questionnaire (UEQ)

Technology Acceptance Model (TAM)

- TAM items elicit **likelihood ratings** rather than agreement.
- The purpose of the model was to **predict future** use of a product rather than rating the experience of its actual use.

Technology Acceptance Model (TAM)

- Perceived Usefulness (PU)

Technology Acceptance Model							
Perceived Usefulness (PU)	Likely						Unlikely
	Extremely	Quite	Slightly	Neither	Slightly	Quite	Extremely
1. Using [this product] in my job would enable me to accomplish tasks more quickly.							
2. Using [this product] would improve my job performance.							
3. Using [this product] in my job would increase my productivity.							
4. Using [this product] would enhance my effectiveness on the job.							
5. Using [this product] would make it easier to do my job.							
6. I would find [this product] useful in my job.							

Technology Acceptance Model (TAM)

- Perceived Ease-of-Use (PEU)

Perceived Ease-of-Use (PEU)	Likely						Unlikely
	Extremely	Quite	Slightly	Neither	Slightly	Quite	Extremely
7. Learning to operate [this product] would be easy for me.							
8. I would find it easy to get [this product] to do what I want it to do.							
9. My interaction with [this product] would be clear and understandable.							
10. I would find [this product] would be clear and understandable.							
11. It would be easy for me to become skillful at using [this product].							
12. I would find [this product] easy to use.							

NASA Task Load Index (TLX)

Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?

Very Low Very High

System Usability Scale (SUS)

Participants are asked to score the following 10 items with one of five responses that range from Strongly Agree to Strongly Disagree:

1. I think that I would **like** to use this system frequently.
2. I found the system unnecessarily **complex**.
3. I thought the system was **easy** to use.
4. I think that I would **need** the **support** of a technical person to be able to use this system.
5. I found the various **functions** in this system were **well integrated**.
6. I thought there was too much **inconsistency** in this system.
7. I would imagine that most people would **learn** to use this system very **quickly**.
8. I found the system very **cumbersome** to use.
9. I felt very **confident** using the system.
10. I needed to **learn a lot** of things **before I could get going** with this system.

Player Experience Inventory (PXI)

Consists of multiple constructs and items, e.g.:

MEANING:

Playing the game was meaningful to me.

The game felt relevant to me.

Playing this game was valuable to me.

MASTERY:

I felt capable while playing the game.

I felt I was good at playing this game.

I felt a sense of mastery playing this game.

IMMERSION:

I was no longer aware of my surroundings while I was playing.

I was immersed in the game.

I was fully focused on the game.

AUTONOMY:

I felt a sense of freedom about how I wanted to play this game.

I felt free to play the game in my own way.

I felt like I had choices regarding how I wanted to play this game.

User Experience Questionnaire (UEQ)

obstructive	o o o o o o o o	supportive
complicated	o o o o o o o o	easy
inefficient	o o o o o o o o	efficient
confusing	o o o o o o o o	clear
boring	o o o o o o o o	exciting
not interesting	o o o o o o o o	interesting
conventional	o o o o o o o o	inventive
usual	o o o o o o o o	leading edge

experiments vs usability testing: which questionnaire would you use?

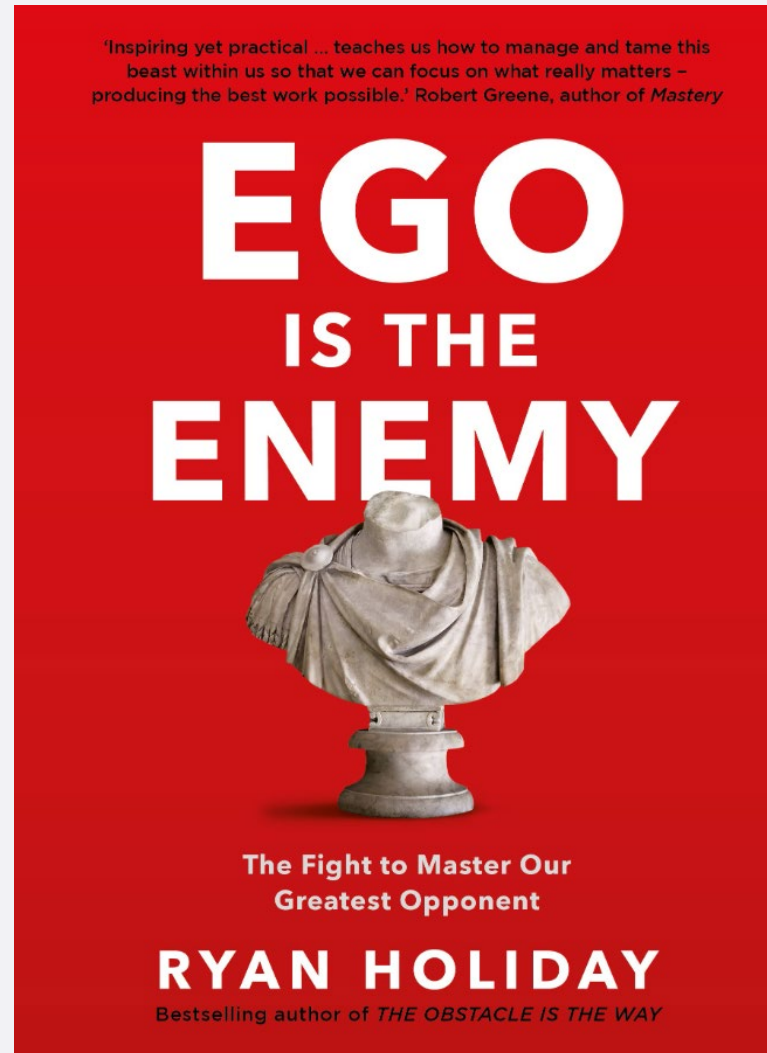
Imagine that you are **testing a new fitness app**. It is the first time for your test users to try an app of this kind, and you want to figure out how they think they will feel about using it over the course of the next few months. Which questionnaire would you use?

- 🧠 Technology Acceptance Model (TAM)
- 🧠 NASA Task Load Index (TLX)
- 🧠 System Usability Scale (SUS)
- 🧠 Player Experience Inventory (PXI)
- 🧠 User Experience Questionnaire (UEQ)

dealing with test users' feedback: the “Ego Problem”

- ⊗ when we design something, we tend to get **attached to** our prototype/**product** – this is quite natural
- ⊗ you might **take the feedback** provided by test users too **personally** and/or get offended by negative feedback
- ⊗ do not get hurt by the feedback that you receive – use it to get better at what you do and to **improve your work!**

dealing with test users' feedback: the “Ego Problem”



field studies



field studies

🌐 where are they done?

- 🌐 anywhere other than a controlled lab – **in natural settings**, the ‘real world’

🌐 what are they?

- 🌐 **observations, interviews and logs** collected in natural settings
- 🌐 ‘in the wild’ is a popular term for HCI evaluation through field studies
- 🌐 aim to understand what users do naturally and how technology is used in the context of their day-to-day lives – **ecological validity**

field studies: data collection

🔗 **observation & interviews** produce a range of data, including:




- 🔗 notes
- 🔗 photos
- 🔗 video/audio
- 🔗 logs

🔗 **self-report** is also used to gather information about user experiences:







- 🔗 diary studies
- 🔗 questionnaires

field studies: observation

how?



-  either **from a distance or closely working**
-  note down observations on a specially designed form
-  video recordings may be appropriate

what to look for (base this on evaluation questions)






-  does the user have **any problems** interacting with the tool?
-  **what tasks** does the user carry out?
-  how does the tool mediate **social interactions**?
-  is the tool **used as expected**?
-  do people look like they're **enjoying the interaction**?
-  are there any **disruptions** to the tasks?

field studies: interviews

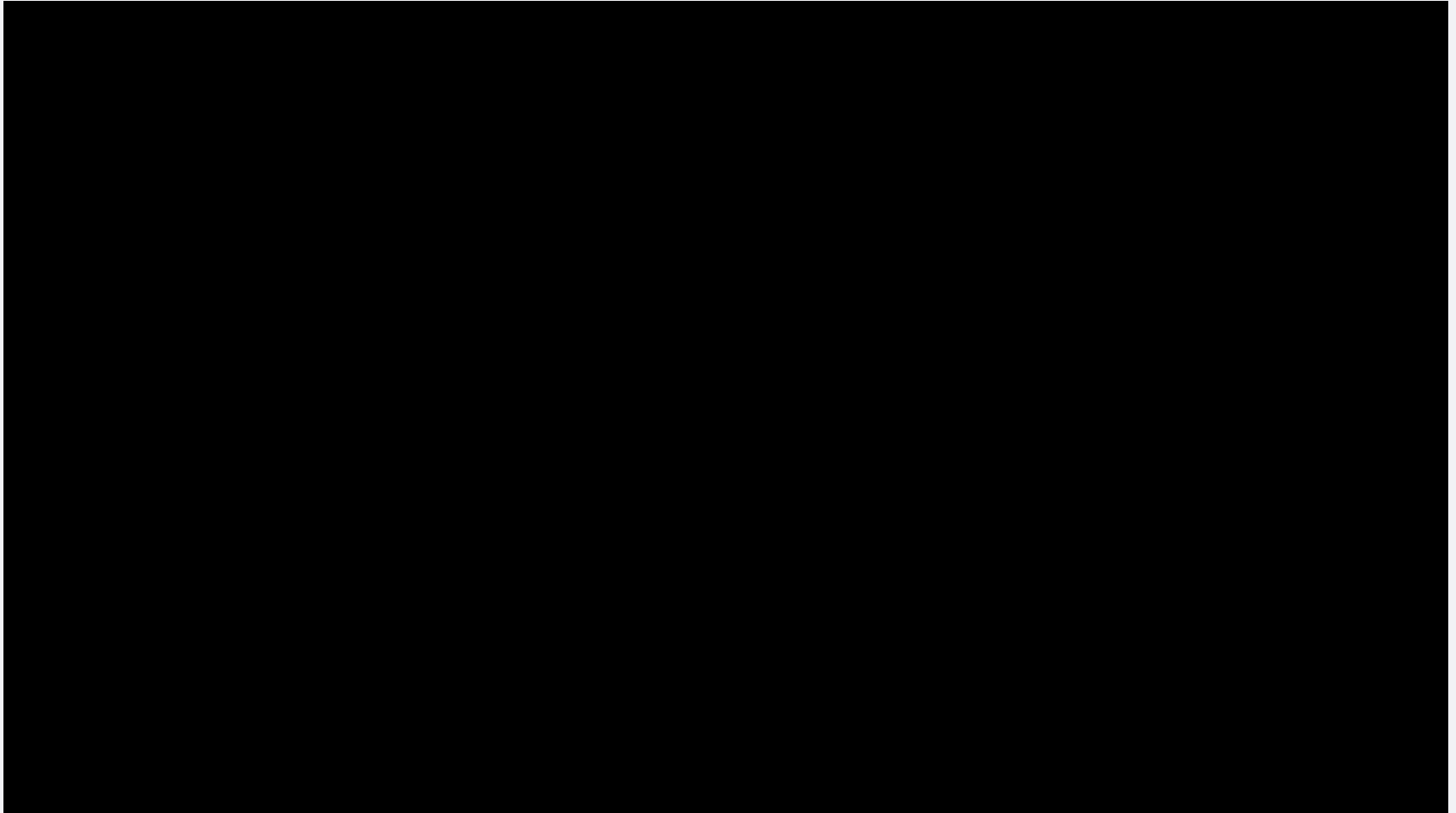
how and when

-  ask questions – either **during or after testing** activities
-  take notes or make audio recordings

what to ask about

-  **satisfaction/enjoyment** of overall experience
-  follow up on interesting observations – ask **why something happened**
-  **clarification** where an event or activity was unclear
-  ask for **examples** to back up general points
-  give the participant a **chance to say anything else** they want

field study: Essence example



field studies: pros & cons

pros

helps understand **what users do naturally** and how technology impacts them in context



evaluating in **real usage** is the only way of finding out whether a design has been successful in its aims



cons



access to settings



lack of control, noise, distractions



hard to capture detail of usability issues



living labs

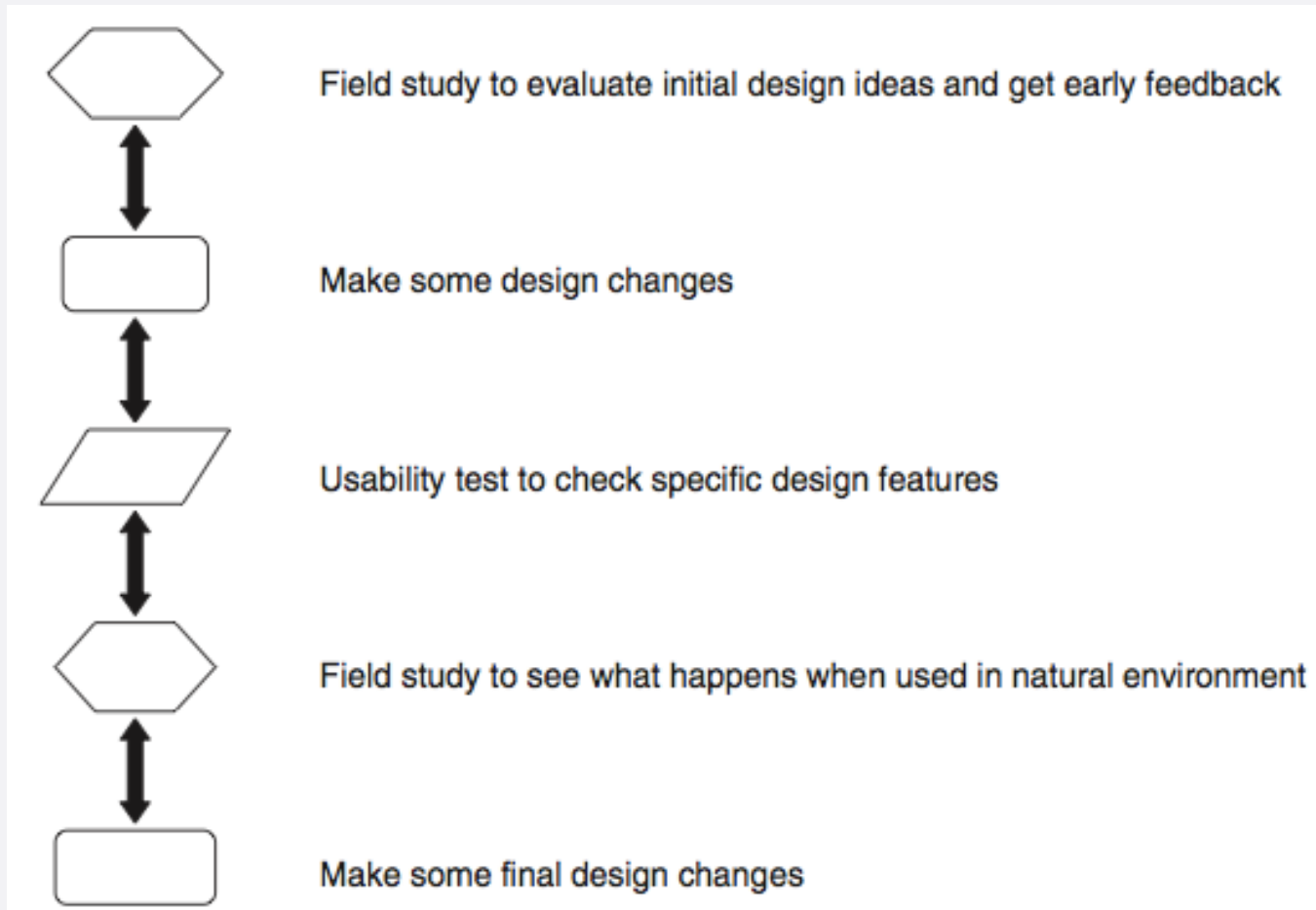


- ⊕ the best of both worlds?
- ⊕ many evaluations are too difficult to do in a usability lab, but field studies are very uncontrolled
- ⊕ people's use of technology in their everyday lives can be evaluated in living labs which aim to emulate real world settings but are wired for data capture
- ⊕ however, these are very costly to set up, and they are still **not** the **real** world

choosing an approach

- ⊕ there is a wide range of approaches to choose from, including those involving users and those which do not require user involvement
- ⊕ every approach has its compromises/ challenges
- ⊕ how do you pick?
- ⊕ often using **two or more methods** which **complement each other** is the best idea

usability testing & field studies can complement each other



evaluation

case study: how would you evaluate this tool?

**MAY CONTAIN CONTENT
INAPPROPRIATE FOR CHILDREN**

Visit www.esrb.org
for rating information

Reminders

 **Drop-in Sessions**

 **Office Hour**

Please check the “Module Contacts” page on Canvas for details!

a bit of **homework** to prepare for
the next seminar: watch this video!



week 7 reading

- Read the “**Introducing Evaluation**” chapter of the Interaction Design book.

