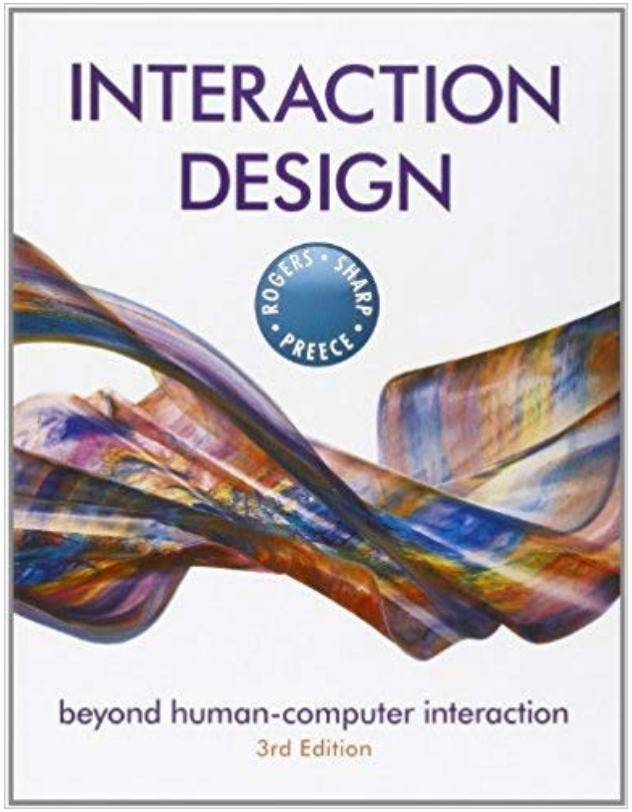
CHALMERS |  UNIVERSITY OF GOTHENBURG

DIT045 H17 Requirements and User Experience

Lecture 12: Usability Testing & Field Studies

Jennifer Horkoff
Email: jenho@chalmers.se



Interaction Design

Evaluation: Usability Testing and Field Studies

Usability Testing - What

- Asking users to perform various tasks with the interface.
  - E.g. add a course, buy a product, post a message, review a timeline etc., etc.
- Observing how they perform:
  - How fast they perform it? *(efficiency)*
  - How often they succeed? *(effectiveness)*
  - Are they confused, lost, dissatisfied etc. at any point?
  - What do they think *afterwards*?
- Analyze Results and revise Interface, e.g.:
  - If it takes too long to locate a button, make it more salient.
  - If user doesn't know what to do at a given point in time, add e.g. constraints.

Usability Testing – How

- Select a number of representative users.
  - Called **participants** or **subjects**.
  - Normally 5-10 users.
- Define a number of typical interaction tasks.
- Invite participants to *the lab.*
- Acquire informed consent to participate.
- Have them perform each of the tasks.
  - Usually < 30 minutes to avoid fatigue effects.
- Take performance measurements.

Usability Testing – How

- In the beginning: May or may not guide the user thought the screens.
  - No, if learnability is something to be measured.
  - Yes, if system is by its nature too complex to be learned without help.
- During: ask users to perform tasks in order to achieve specified goals.
  - “Please find course ITEC3230 Fall 2018 and enrol it.”
  - “Pick-up the Atlas MP3 Player 5GB, add it to cart and go to check-out.”
  - “Find who has responded to your post.”
  - “Find which photos of yours have been tagged.”
- During: measure usability indicators (next slide)

Usability Testing – What to Measure (1/2)

- Quantitative Data:
  - Time to complete a task.
    - Measure of *efficiency.*
  - Time to complete a task after a specified time away from the product.
    - Measure of *memorability.*
  - Number and type of errors per task.
    - Measure of *safety or effectiveness*.
  - Number of errors per unit of time.
    - Measure of *safety or effectiveness*.
  - Number of navigations to online help or manuals.
    - Measure of *learnability*.
  - Time to figure out what to do.
    - Measure of *learnability*.
  - Number of users making a particular error.
    - Measure of *safety or effectiveness*.
  - Number of users completing task successfully.
    - Measure of *safety or effectiveness*.

Usability Testing – What to Measure (2/2)

- Qualitative Data:
  - Body language and facial expressions.
    - Expressions of frustration, dissatisfaction, confusion, hesitation.
  - Articulated comments
    - User is probably talking while using the system, expressing dissatisfaction, difficulty, confusion, lack of understanding.
- Use video camera to record both the face of the participants and the interface he/she is using.
- Voice-record user's comments while performing the tasks.
- Must have acquired written-signed consent beforehand to do all that!

Usability Testing – After the Tasks

- Administer **questionnaire** to assess other parameters.
  - Overall experience and satisfaction level.
  - Perceived ease of use/intuitiveness/quality.
  - Suggestions for improvement.
  - Priorities for addressing issues.
  - …
- Questions can be
  - closed-ended (= quantitative analysis)
  - open-ended (= qualitative analysis)
- Use a standard usability assessment questionnaire.
  - SUS, SUMI, PSSUQ, UMUX

The System Usability Scale (SUS)

- Very popular in the industry.
  - About 43% usage.
- Simple and quick.
- Two versions:
  - Standard[1] (alternating positive and negative-tone questions)
  - Positive[2,3] (only positive).
- 10 Likert style questions
  - Responses from 1 (strongly disagree) to 5 (strongly agree)
  - For neutral it is 3.
- Overall SUS score calculation:
  - Positive items (for standard and positive format): score – 1.
  - Negative items (for standard format): 5 – score.

[1] J. Brooke, A 'quick and dirty' usability scale. Usability Evaluation in Industry (pp.189-194). 1996
[2] J. Sauro and J. R. Lewis, When designing usability questionnaires does it hurt to be positive? CHI 2011
[3] J. Sauro and J. R. Lewis, Quantifying user experience: Practical Statistics for User Research, Morgan Kaufman, 2012.

The System Usability Scale (SUS)

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 to support of a technical person to be able to use this system.
5	I found that 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.
8	I felt very confident using the system.
10	I needed to learn a lot of things before I could get going with this system.

SUS Scoring

	Question	Response	Score
1	I think that I would like to use this system frequently.	Agree (4)	3 (4-1)
2	I found the system unnecessarily complex.	Disagree (2)	3 (5-2)
3	I thought the system was easy to use.	Disagree (2)	1 (2-1)
4	I think that I would need to support of a technical person to be able to use this system.	Strongly Agree (5)	0 (5-5)
5	I found that the various functions in this system were well integrated.	Agree (4)	3 (4-1)
6	I thought there was too much inconsistency in this system.	Agree (4)	1 (5-4)
7	I would imagine that most people would learn to use this system very quickly.	Strongly Disagree (1)	0 (1-1)
8	I found the system very cumbersome to use.	Strongly Disagree (1)	4 (5-1)
8	I felt very confident using the system.	Agree (4)	3 (4-1)
10	I needed to learn a lot of things before I could get going with this system.	Neutral (3)	2 (3-1)
	TOTAL		20 * 2.5 = 50%

SUS Curved Scale

SUS Score Range	Grade	Percentile Range
84.1 – 100	A+	96 – 100
80.8 – 84.0	A	90 – 95
78.9 – 80.7	A-	85 – 89
77.2 – 78.8	B+	80 – 84
74.1 – 77.1	B	70 – 79
72.6 – 74.0	B-	65 – 69
71.1 – 72.5	C+	60 – 64
65.0 – 71.0	C	41 – 59
62.7 – 64.9	C-	35 – 40
51.7 – 62.6	D	15 – 34
00.0 – 51.6	F	0 –14

[1] J. Sauro and J. R. Lewis, Quantifying user experience: Practical Statistics for User Research, Morgan Kaufman, 2012.

Usability Testing – Summary of Data Collected

- Performance data during the task.
  - Logged by the interface (e.g. clicks, keypresses)
  - Video-recorded and processed afterwards.
- Experience data during the task.
  - Voice/video recordings of participant and the interface.
- Experience data after the task.
  - Questionnaire and/or discussion with the analyst in the end.

Usability Testing – Data Analysis

- Performance data during the task.
  - Identify possible efficiency, safety, effectiveness, learnability, memorability issues.
- Experience data during the task.
  - Explain/identify/contextualize performance issues.
  - Assess user experience.
- Experience data after the task.
  - Assess user experience.
  - Elicit attitudes.
  - Elicit priorities.

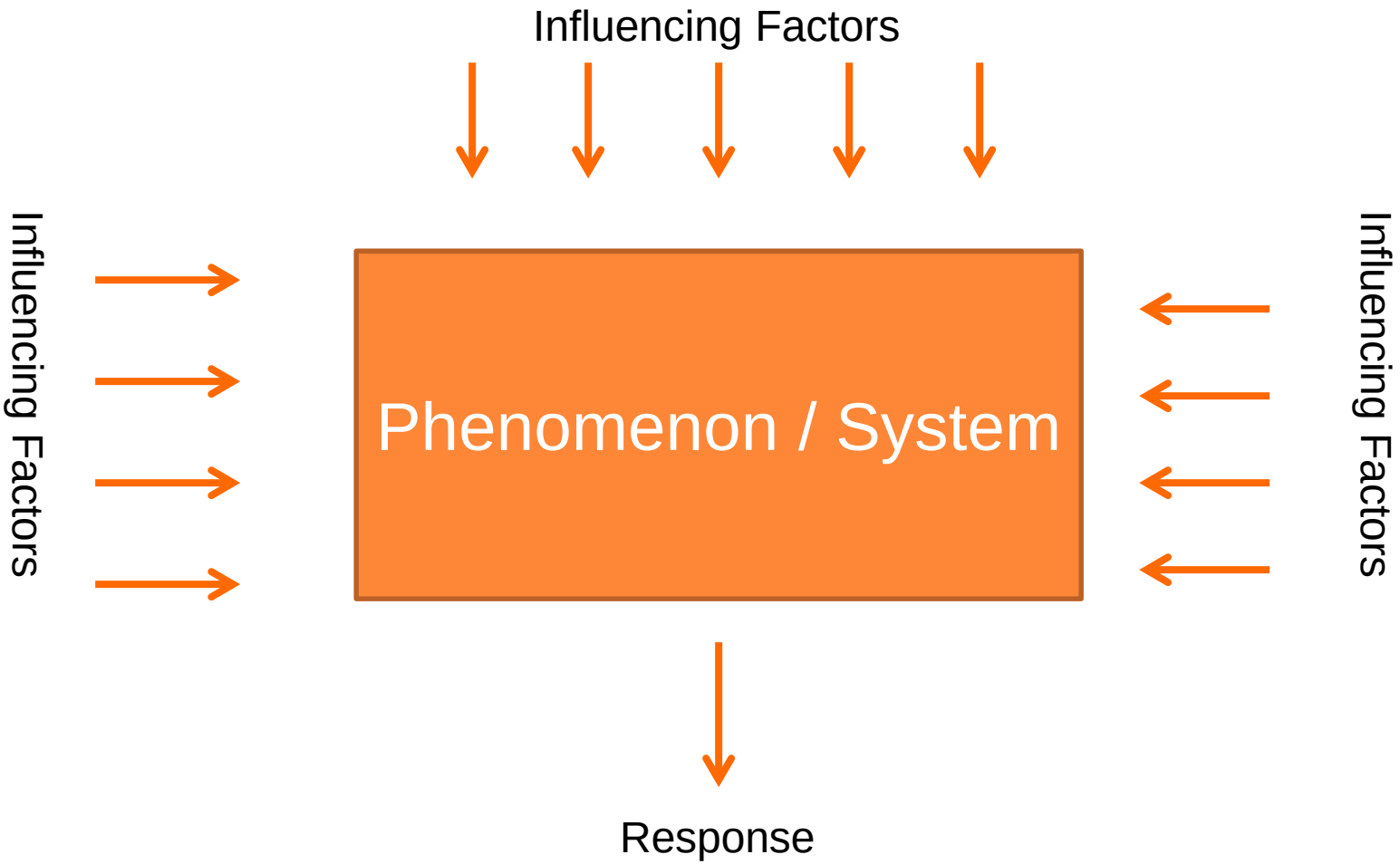
Usability lab with observers watching a user & assistant

Portable Equipment for Use in the Field

Usability Experiments

- When you have to compare two candidate interfaces.
  - E.g. an old and a new one.
- There is a need for generalization for a larger class of systems (e.g. interested in statistical significance)
- Larger and representative sample sizes.
- More rigorous control of other influencing factors.

What are experiments?



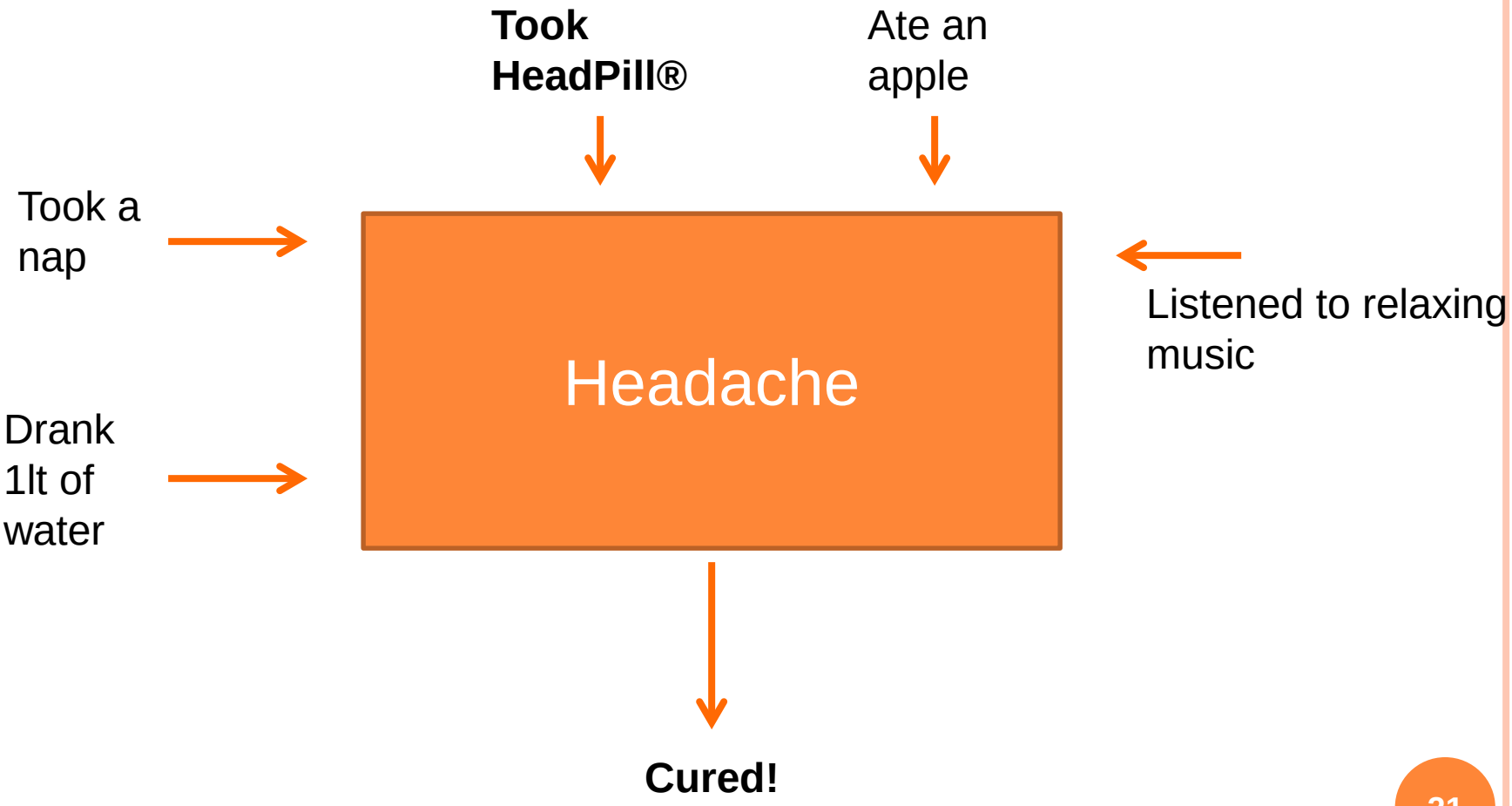
Experiments - Goal

- Establish whether the **response** can be actually **attributed** to one of the possible influencing factors.
  - I.e. say: “it was because of ABC factor that we observed XYZ response.”
- In usability experiments we want to say:
  - “it was because of the interface design that a user took less time to complete a task (and not because e.g. the user was smarter)”

Experiments - Example

- Assume that I produce the HeadPill® , which, **I claim**, cures headache.
  - Would you believe me and go buy the HeadPill?
  - What proof do you need in order to believe me?
- Assume that I give the pill to a participant.
- The participant comes back after a few hours and says he was *cured*!
- Are you convinced?

Experiments - Example



Experiments - Example

- The pill ***may*** have cured the headache…
- … **but** it may have also been **the apple** that cured the headache…
- … or the **1 litre of water** that the participant drank…
- … or the fact that he **took a nap**.

- So: we cannot attribute the response (that the headache was cured) to the pill.
- We need to **rule out all** other factors.
  - How??
  - By **controlling for** them.

Experiments - Technique

- Take 2 groups of participants A and B, all with headache.
- Participants in A eat an apple, drink 1lt of water, take a nap, listen to relaxed music **and take the pill.**
- Participants in B eat an apple, drink 1lt of water, take a nap, listen to relaxed music **but *do not* take the pill.**
- I.e. you **control for** apples, water, sleep, relaxed music.
- If A are cured and B are not cured:
  - You know it **must have been** the pill.
  - All else was equal for A and B! There is no other explanation. (*)

Experiments - Terminology

- **Independent Variable.**
  - The influencing factor you are tweaking in order to see if it has an effect to the **dependent variable**.
  - It is up to you to tweak it, that's why it is "independent".
  - **Example:** whether or not to take the pill.
- **Dependent Variable.**
  - The response/effect you are wondering if it should be attributed to the independent variable.
  - **Example:** whether headache goes away.
- **Confounding factors.**
  - Other influencing factors that, if left uncontrolled, they won't let you prove the connection.
  - **Example:** eating the apples, drinking the water, sleeping, etc.
- **Experimental Condition.**
  - The state of the independent variable at a given phase of the experiment.

Usability Experiments

- Same logic like any experiment.
- Used to **compare** two interfaces A (e.g. an existing) and B (a redesigned one).
- Have some subjects use A and some subjects use B.
- Keep all else equal:
  - The time of the day, the amount of sleep subjects had, the environment, the keyboard/mouse, the age/skill of the subjects.
- Observe: do subjects who use B perform better (e.g. faster) than those who used A?
  - If yes, and everything else is held equal (= is controlled for) then it is probably because of B being a better design that we observe the better performance.

Usability Experiments

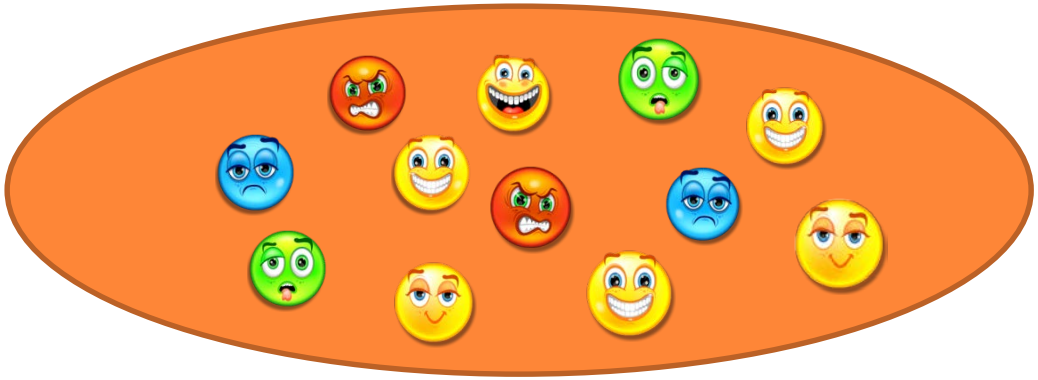
- **Dependent Variable:**
  - How long does it take to perform *a given* task (e.g. enrol a course?)
- **Independent Variable**
  - Use **System A** or **System B**
- **Confounding factors**
  - The time of the day, the amount of sleep participants had, the environment, etc.
- **Experimental Condition.**
  - Whether the given task is performed using System A or System B.

Samples and Individual Differences

- You want to test if a coin is fair.
- How many times do you toss it?
  - One time?
  - Two times?
  - More times?
- Clearly the more times the better. A bigger sample (of coin tosses) is more “representative” of the population of all possible coin tosses.
- The same logic applies to participants.
- Participants have different characteristics and capabilities:
  - Visual and motor skills, cognitive skills, attention levels, culture, background, education…
  - They have slept different amounts of time, eaten or not before they came to your experiment, may or may not be having a bad day, may or may not be stressed/distracted about something etc…
- Solution: choose many of them to level out these **individual differences.**

Assigning Participants to Experiments

- When comparing interface A with interface B we need to decide which participants will use A and which participants will use B.



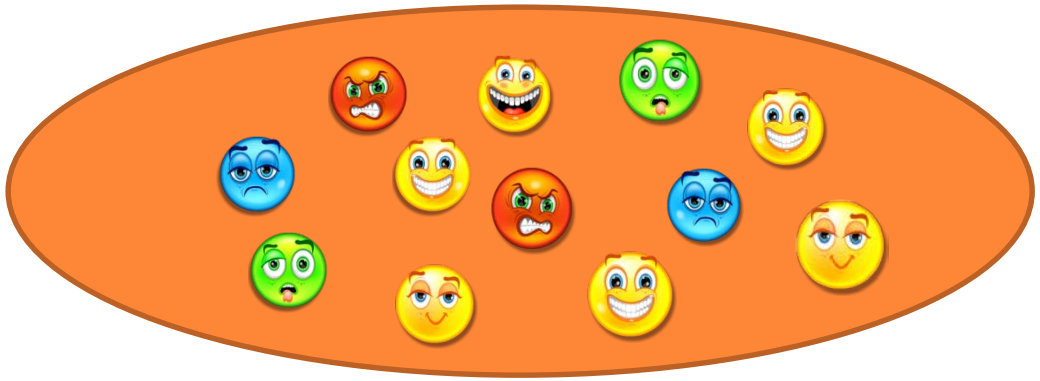
Interface A

Interface B

Experimental designs

- **Different participants** (“between subjects”) - single group of participants is allocated randomly to the experimental conditions.
- **Same participants** (“within subjects”) - all participants appear in both conditions. (**counterbalancing** to neutralize the learning effect)
- **Matched participants** - participants are matched in pairs, e.g., based on expertise, gender, etc.

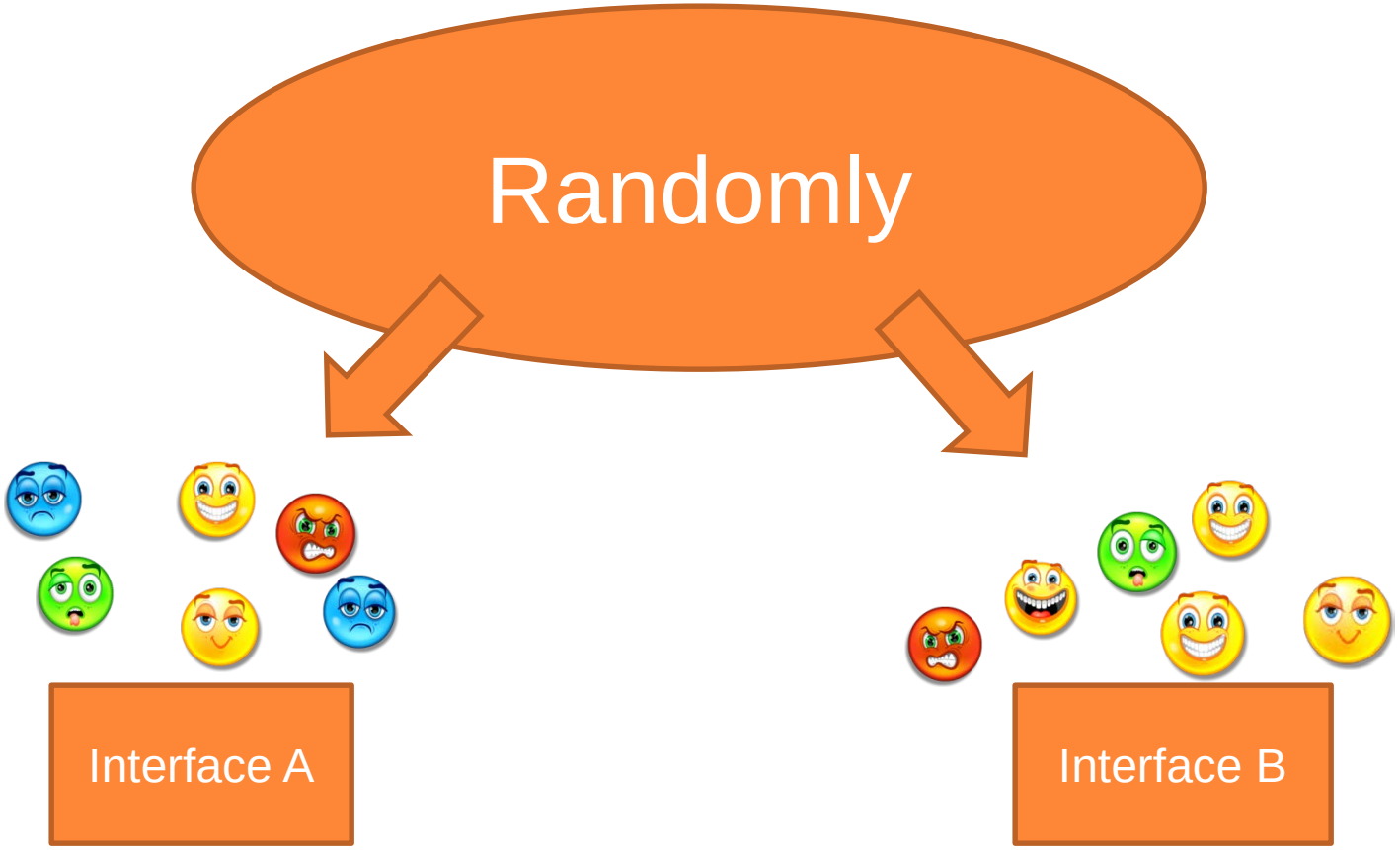
Different Participants – Between Subjects



Interface A

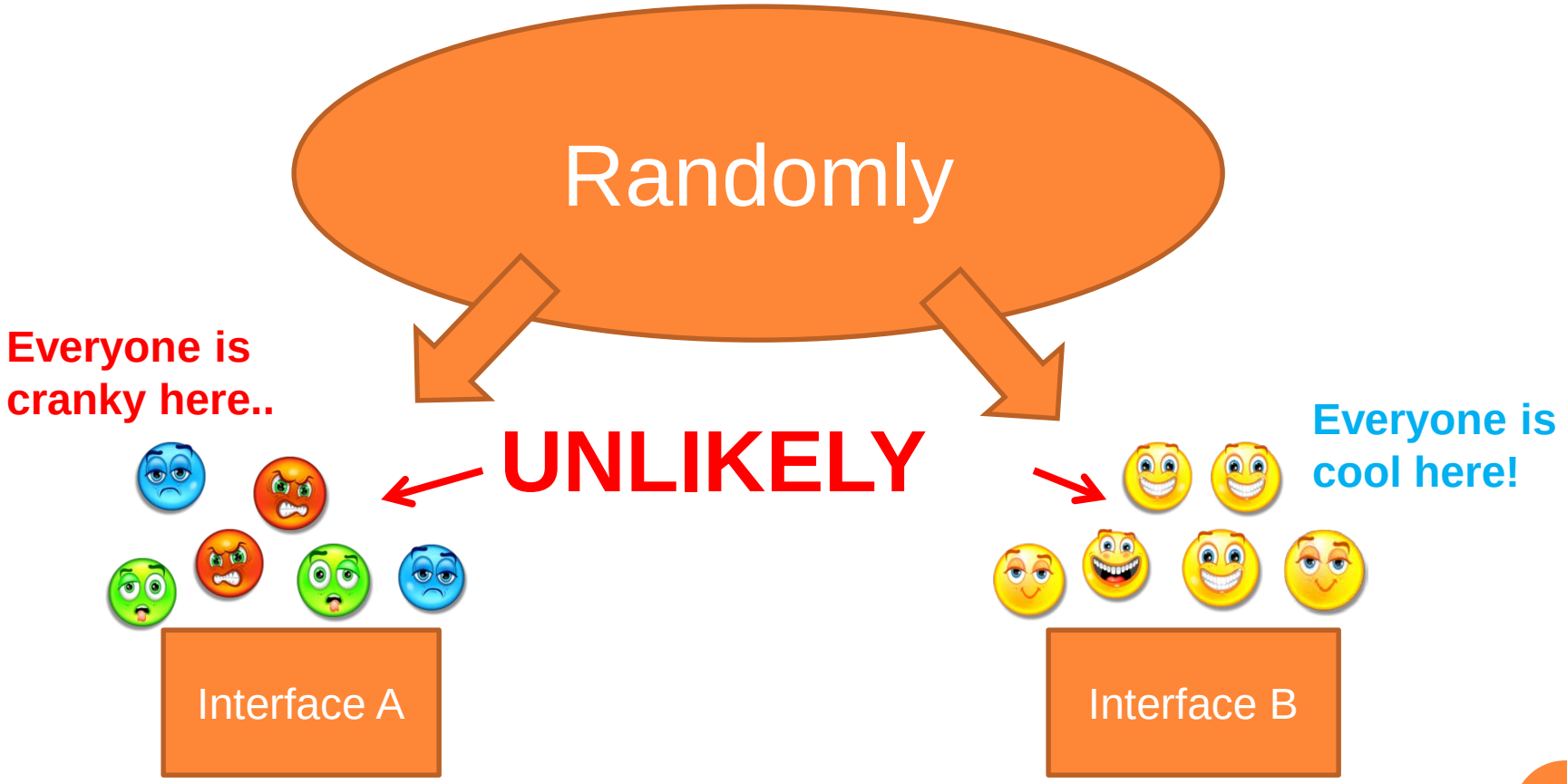
Interface B

Different Participants – Between Subjects

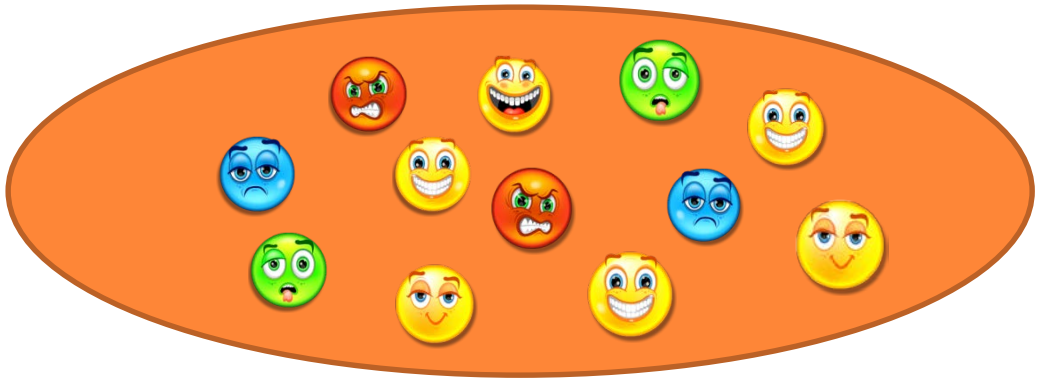


Through randomly distributing to conditions, when the sample is large enough, individual differences are unlikely to group together.

Different Participants – Between Subjects



Same Participants – Within Subjects



Interface A

Interface B

Same Participants – Within Subjects

Stage 1:
everybody uses
interface A

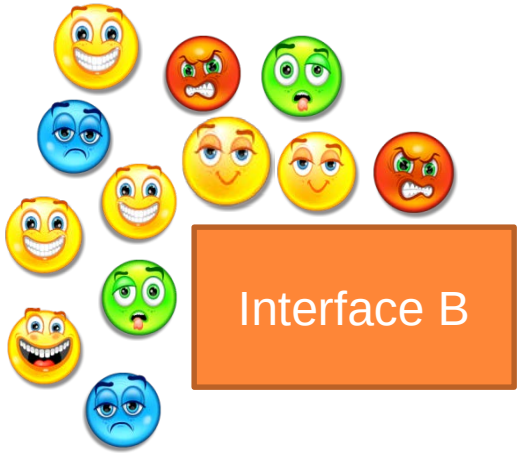
Interface A

Interface B

Same Participants – Within Subjects

Stage 2:
everybody uses
interface B

Interface A

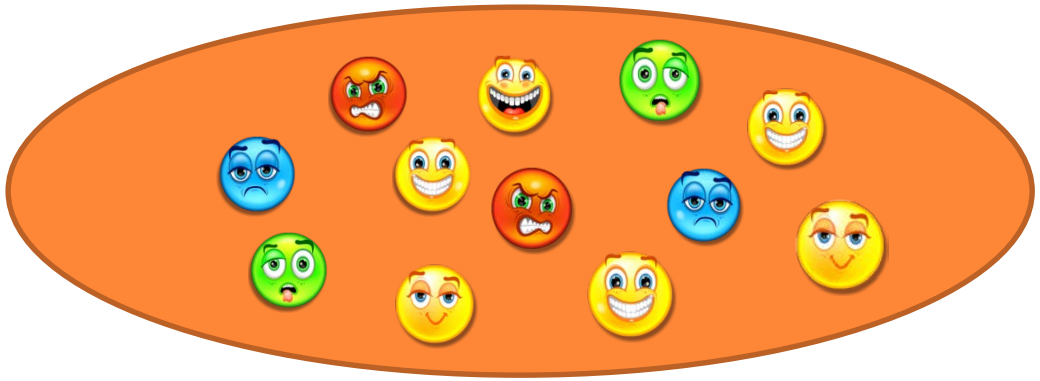


Interface B

Same Participants – Within Subjects: ISSUE

- **Learning Effect / Order effect**
  - When participants use interface A they familiarize themselves with it and the experimental process.
  - They are more likely to perform better using Interface B just because of this.
  - Thus ordering becomes a **confounding factor.**
  - Differences in the responses may not be due to the quality of the interface but the sequence in which it was tested.
- Solution:
  - Counterbalancing.
  - Split participants: half do A → B, and half do B → A.
  - Learning effects **still exist** but **cancel out**.
    - (Assuming they are symmetric)

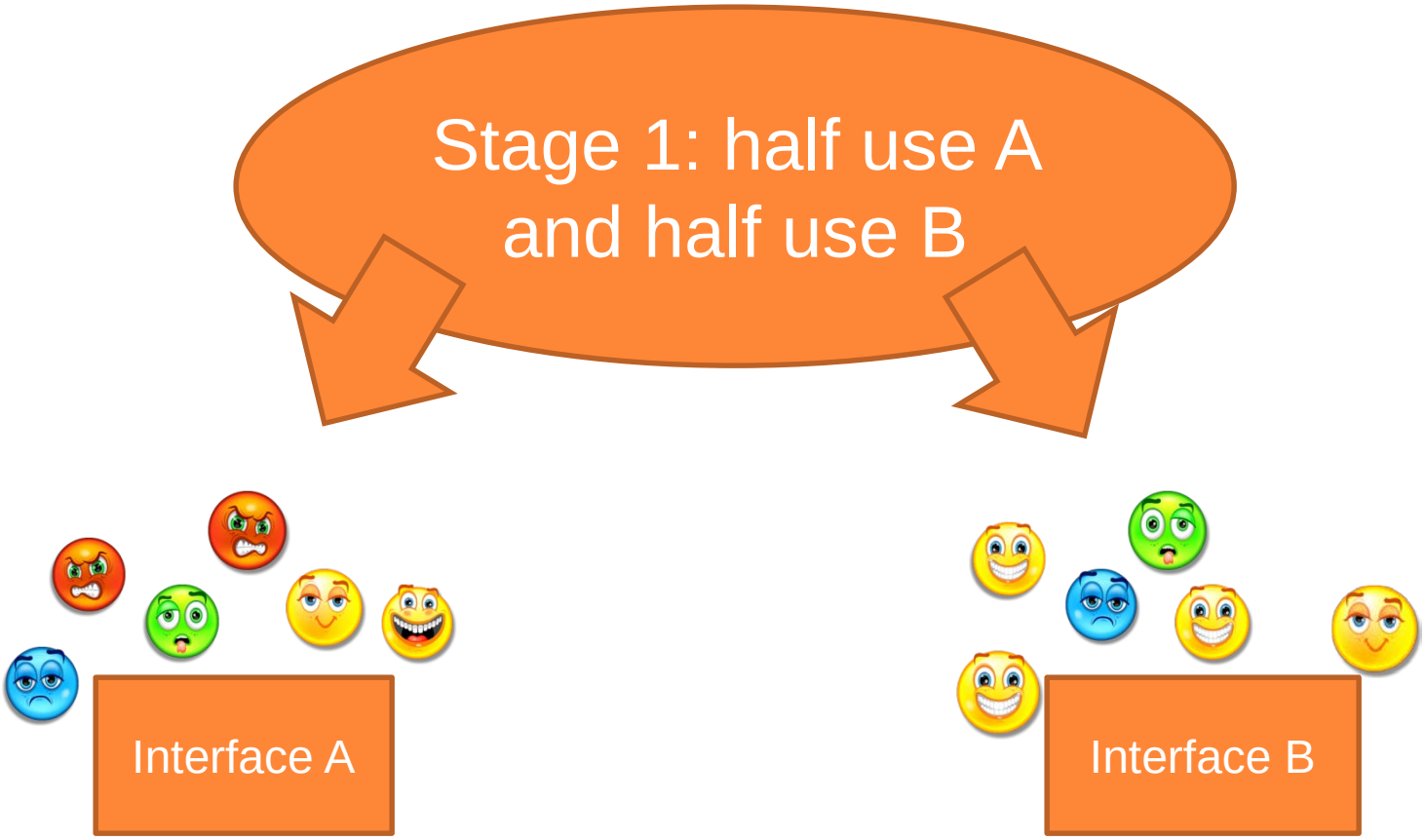
Within Subjects with Counter balancing



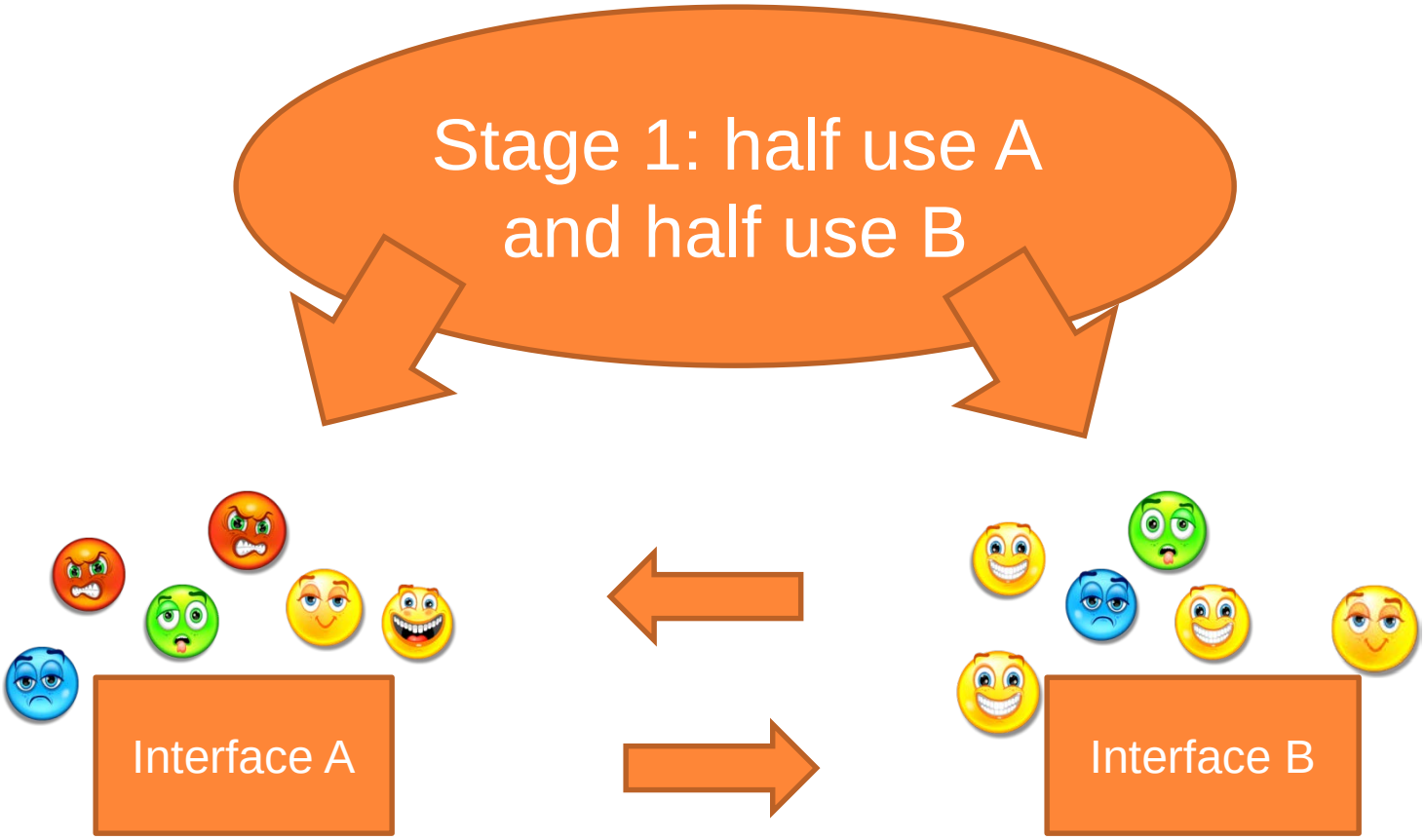
Interface A

Interface B

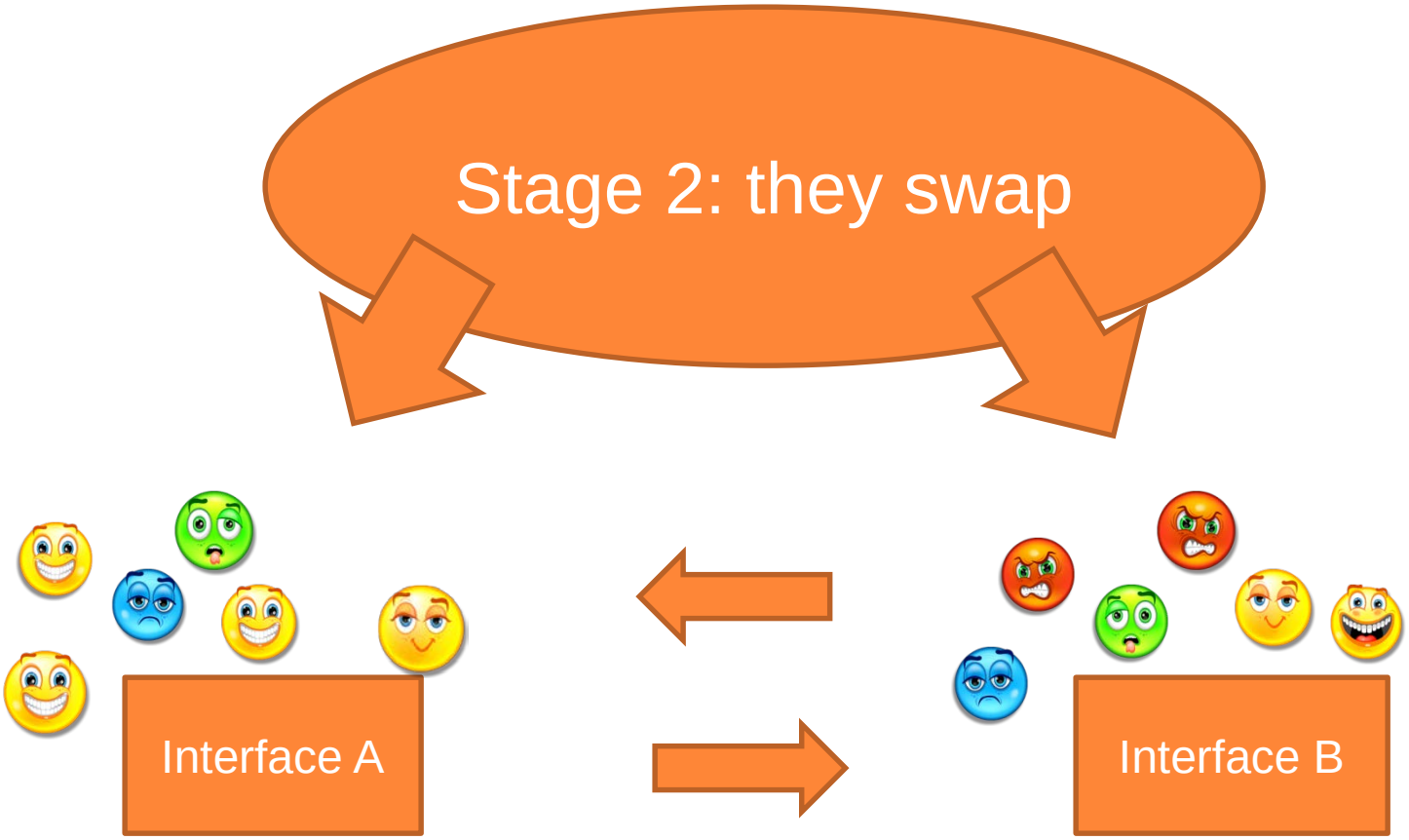
Within Subjects with Counter balancing



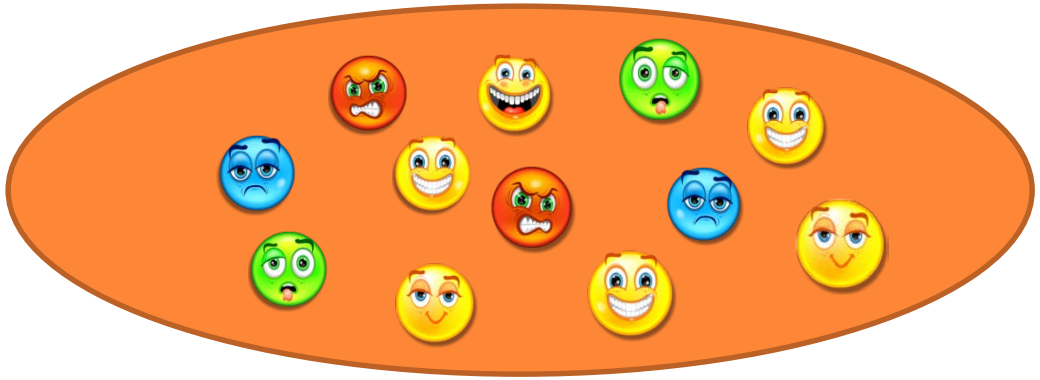
Within Subjects with Counter balancing



Within Subjects with Counter balancing



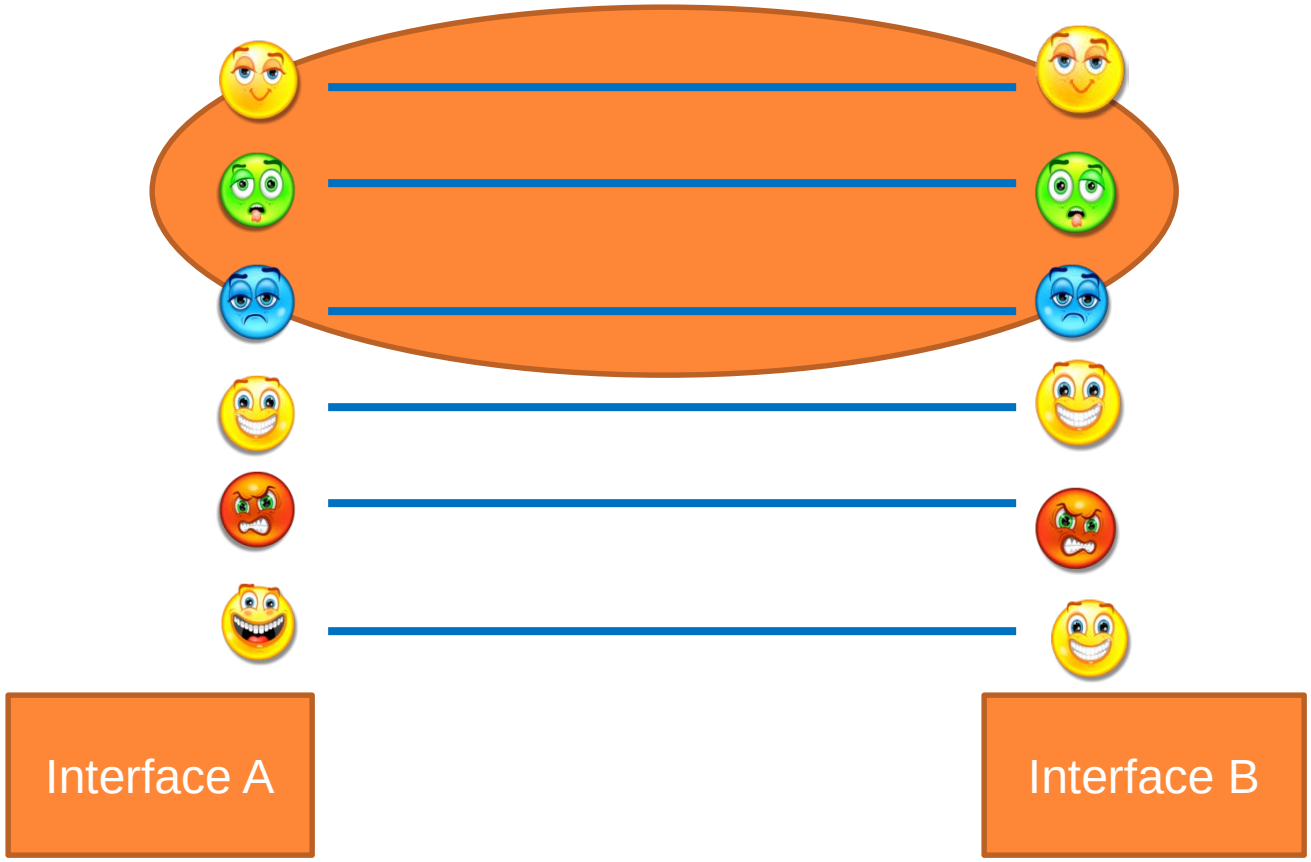
Matched Participants



Interface A

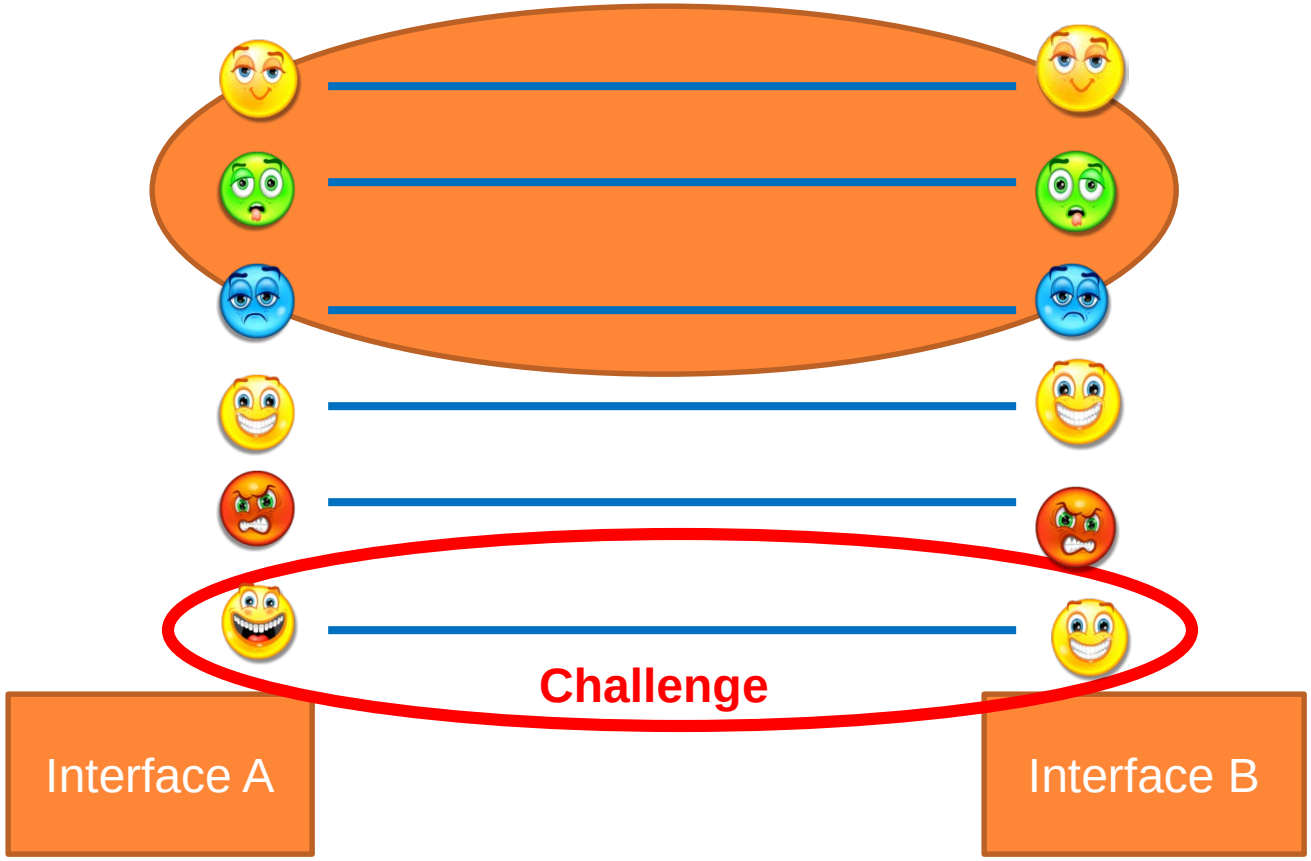
Interface B

Matched Participants



Like between subjects, but not leave it to randomness.

Matched Participants



Like between subjects, but not leave it to randomness. That might be challenging.

Usability Experiment - Example

- You have two course enrolment interfaces: the Old one and the New one.
- You claim that the New one allows users to enrol courses much faster than the Old one.
- Design an experimental study that could support or refute your claim.
- Specify:
  - Independent Variables
  - Dependent Variables
  - Nuisance Variables / Confounding Factors
  - Experimental Conditions (what participants actually do)
- How do you split participants for:
  - Between Subjects?
  - Within Subjects?
  - Matched Participants?

Different, Same, Matched Participant Design

Design	Advantages	Disadvantages
Different	No order effects	Many subjects & individual differences a problem
Same	Few individuals, no individual differences	Counter-balancing needed because of ordering effects
Matched	Same as different participants but individual differences reduced	Cannot be sure of perfect matching on all differences

Usability Testing vs. Research

Usability testing

- Improve products
- Few participants
- Results inform design
- Usually not completely replicable
- Conditions controlled as much as possible
- Procedure planned
- Results reported to developers

Experiments for research

- Discover knowledge
- Many participants
- Results validated statistically
- Must be replicable
- Strongly controlled conditions
- Experimental design
- Scientific reported to scientific community

Field Studies

- Field studies are done in natural settings.
- The aim is to understand what users do naturally and how technology impacts them.
- Field studies can be used in product design to:
  - identify opportunities for new technology;
  - determine design requirements;
  - decide how best to introduce new technology;
  - evaluate technology in use.

Field Studies

- Good for understanding **appropriation**:
  - Understanding how users, integrate and adopt technology to their needs, desires and culture.
- Data collection:
  - Notes, pictures, recordings
  - Video
  - Logging (and often prompting)
- Analysis:
  - Qualitative analysis of various types. (e.g. activity theory)

Key Points

- Testing is a central part of usability testing.
- Usability testing is done in controlled conditions.
- Usability testing is an adapted form of experimentation.
- Experiments aim to test hypotheses by manipulating certain variables while keeping others constant.
- The experimenter controls the independent variable(s) but not the dependent variable(s).
- There are three types of experimental design: different-participants, same- participants, & matched participants.
- Field studies are done in natural environments.
- Typically observation and interviews are used to collect field studies data