

Topic 8: From Controlled to Natural Settings (Part 2)

SECV2113 Human-Computer Interaction

Faculty of Computing

Universiti Teknologi Malaysia



01 USABILITY TESTING

02 CONDUCTING EXPERIMENTS

03 IN-THE-WILD STUDIES

USABILITY TESTING

....

What is Usability Testing?

- Involves testing how people perform on products in controlled settings
- The people being tested are observed and the time it takes them to complete a task and the number and kinds of errors they make are recorded
- Data is recorded on video and key presses are logged
- The data is used to calculate performance times and to identify and explain errors
- User satisfaction is evaluated using questionnaires and interviews
- Observations about how the product is used in more natural contexts, including in-the wild may be included

Quantitative Measures

- Number of **participants** successfully **completing the task**
- **Time** to complete the task
- **Time** to complete the task after **time away** from it
- Number and **type of errors per task**
- Number of **errors per unit of time**
- Number of times **people navigate to an item** such as online help
- Number of **people making the same or similar errors**

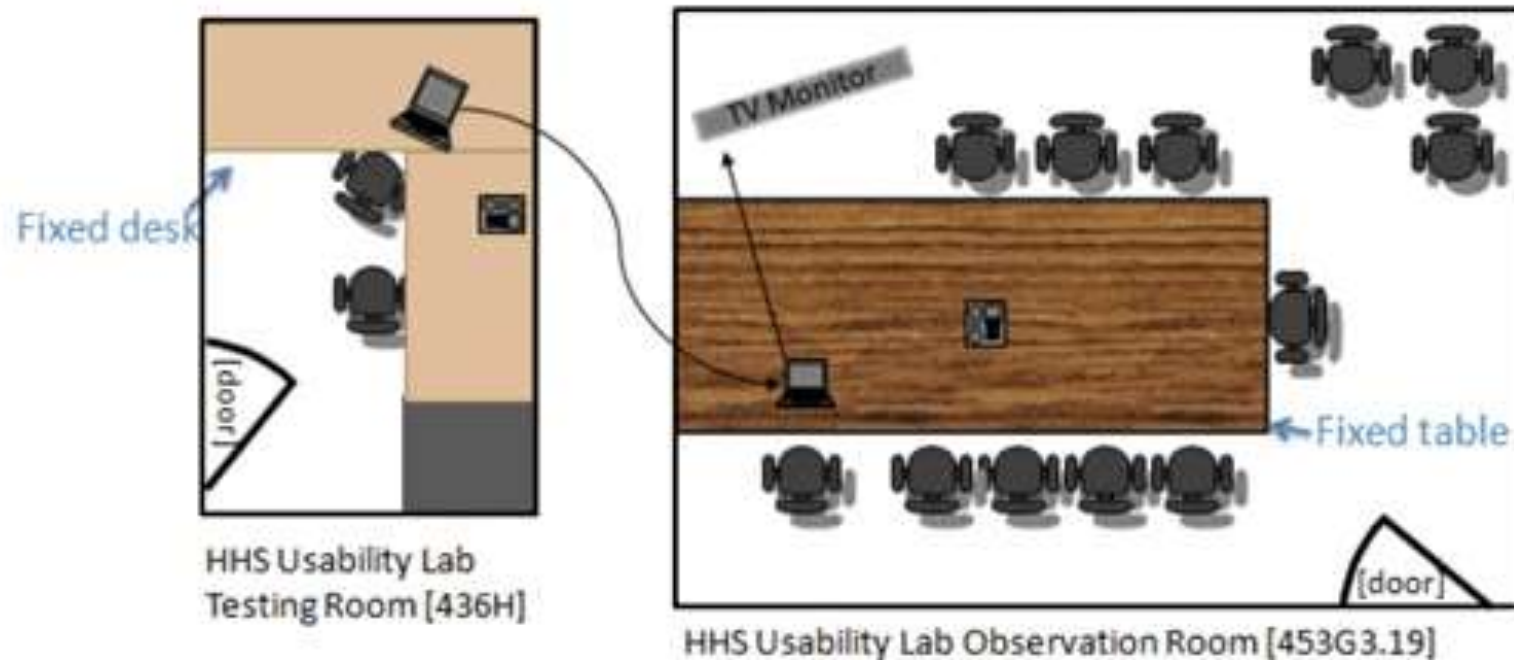
Source: Wixon and Wilson, 1997

Usability lab with designers watching a user and assistant through a one-way mirror and recording



MIMOS UX Lab

Layout of a usability lab used by US Health & Human Services Dept.



source: <https://www.usability.gov/how-to-and-tools/guidance/hhs-usability-lab.html>

Tobii Glasses Mobile Eye-Tracking System



Source: <https://www.tobiipro.com/news-events/on-demand-webinars/Wearable-eye-tracking-for-research/>



Source: Dalton et al., 2015, p.3891. Reproduced with permission of [ACM Publications](#).

Equipment is Getting Smaller

- As more testing **is done remotely and in-the-wild**, equipment has got smaller
- New lightweight models of **eye-tracking glasses** that look like ordinary glasses
- **Video and audio recording** is often done **using mobile phones** which produce increasingly good quality with each new model of phone
- **Replace bulky recording devices and tripods**, also more discrete

Remote Testing

- **Participants** can **carry out tasks in their own environment** without an evaluator being present
- Zoom, Teams and other digital communications are often used
- Testing can be **synchronous or asynchronous**
- Advantage of remote testing, especially asynchronous testing, is that several participants can **be tested at the same time in their own environments**
- Covid-19 pandemic encouraged creative ways of conducting remote testing that was safe
- Remote testing may **make it easier for people with disabilities to participate**
- Downside: no one present to help with **equipment problems** and **less personal** than co-present testing

Case 1: iPad usability

(Budiu & Nielsen, 2010)

- Now a classic study – innovative in its day!
- Study conducted in two cities: Fremont, CA and Chicago, IL
- Tests had to be done quickly, as information was needed by third-party app developers
- Also needed to be done secretly so that the competition was not aware of the study before the iPad was launched
- Seven participants with over three months experience with iPhones

iPad Usability Testing Procedure

- Signed an **informed consent form** explaining:
 - What the **participant** would **be asked to do**
 - The **length of time** needed for the study
 - The **compensation** that would be offered for participating
 - Participants' right to **withdraw from the study** at any time
 - A **promise** that the **person's identity** would **not be disclosed**
 - An **agreement** that the data collected would **be confidential** and available **only to the evaluators**
- Participants were asked to explore the iPad
- Next, they were asked to perform randomly-assigned specified tasks

Examples of the Tasks used in the iPad Evaluation

App or website	Task
iBook	Download a free copy of <i>Alice's Adventures in Wonderland</i> and read through the first few pages.
Craigslist	Find some free mulch for your garden.
eBay	You want to buy a new iPad on eBay. Find one that you could buy from a reputable seller.
<i>Time</i> Magazine	Browse through the magazine and find the best pictures of the week.
Epicurious	You want to make an apple pie for tonight. Find a recipe and see what you need to buy in order to prepare it.
Kayak	You are planning a trip to Death Valley in May this year. Find a hotel located in the park or close to the park.

Adapted from Budiu and Nielsen, 2010

Source: iPad App and Website Usability Study. Used courtesy of the [Nielsen Norman Group](#).

Problems and Actions

- Examples of problems detected:
 - Accessing the Web was difficult
 - Lack of affordance and feedback
 - Getting lost in an application
 - Knowing where to tap
- Actions by evaluators:
 - Reported to developers
 - Made available to public on Nielsen Norman Group
- Accessibility for all users is important
- Study did not address how iPad would be used in people's everyday lives
- Another study was done a year later to examine this and other issues that there was insufficient time to address in this study

Case 2: Remote testing with extended VR (XR)(Sanni Siltanen et al., 2021)

- Study at **Krone Corp. (Finland)** to examine **new immersive experiences in VR**
- Needed to find ways **to test experts and company employees remotely** because of Covid-19 pandemic
- Also needed **a sophisticated setup** to test the collaborative XR platforms across multiple locations
- Testing was conducted **remotely with experts** in 8 countries: Finland, India, China, Germany, Indonesia, Malaysia, USA, United Arab Emirates
- Some tests in **participants premises** where special disinfecting routines had to be carried out, **other tests were done remotely**
- Little advice available about how to do remote testing during a pandemic

A Participant in a Test Session



Source: Siltanen et al. (2021) MDPI / CC BY 4.0

Screenshots from the DesignSpace VR Environments



Source: Siltanen et al. (2021) MDPI / CC BY 4.0

Testing setup used with DesignSpace environments



Source: Siltanen et al. (2021) MDPI / CC BY 4.0

CONDUCTING EXPERIMENT S

....

Experiments

- Test hypotheses
- Predict the relationship between two or more variables
- Independent variable is manipulated by the researcher
- Dependent variable influenced by the independent variable
- Typical experimental designs have one or two independent variables
- Validated statistically and replicable

Experimental Designs

- **Different participants** (between participants):
 - Single group of participants is allocated randomly to the experimental conditions
- **Same participants** (within participants):
 - All participants participate in both conditions
- **Matched participants** (pairwise design):
 - Participants are matched in pairs, for example, based on expertise, gender, and so on

Different, Same, Matched Participant Design

Design	Advantages	Disadvantages
Different	No order effects	Many subjects and 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

IN-THE-WILD STUDIES

....

In-the-Wild Studies

- Done in **natural settings**
- **“In-the-wild”** is a term for prototypes being used freely in natural settings – broader and typically **less controlled** than traditional field studies
- Seek to understand **what users do naturally** and **how technology impacts them**
- In-the-wild studies are used in **product design** to:
 - Identify **opportunities** for new technology
 - Determine **design requirements**
 - Decide how best to **introduce new technology**
 - **Evaluate technology** in use

An In-the-wild Study of a Pain-monitoring Device

- Monitoring patients' pain is a known challenge for physicians
- **Goal of the study** was to evaluate the use of a pain-monitoring device for use after ambulatory surgery
- Painpad is a keypad device
- It was usability tested extensively in the lab before being brought into two hospitals
- **Goal** was to understand how Painpad was used in the natural environment and as part of routines in two UK hospitals.
- How pain-monitoring changed with Painpad

Painpad



A tangible device for inpatient self-logging of pain

Source: Price et al., 2018. Reproduced with permission of [ACM Publications](#).

Data Collection and Participants

- Two studies in two hospitals involving 54 people
- 13 males, 41 females
- Privacy was an important concern
- Hospital stay ranged from 1-7 days, mean and median age 64.6, 64.5 years
- Patients given Painpad after surgery and prompted to report pain levels every two hours
- Nurses also collected scores
- All data entered into charts
- Patients in one hospital were given a user-satisfaction survey when they left
- Also rated Painpad on a 1-5 Likert scale

Data Analysis and Presentation

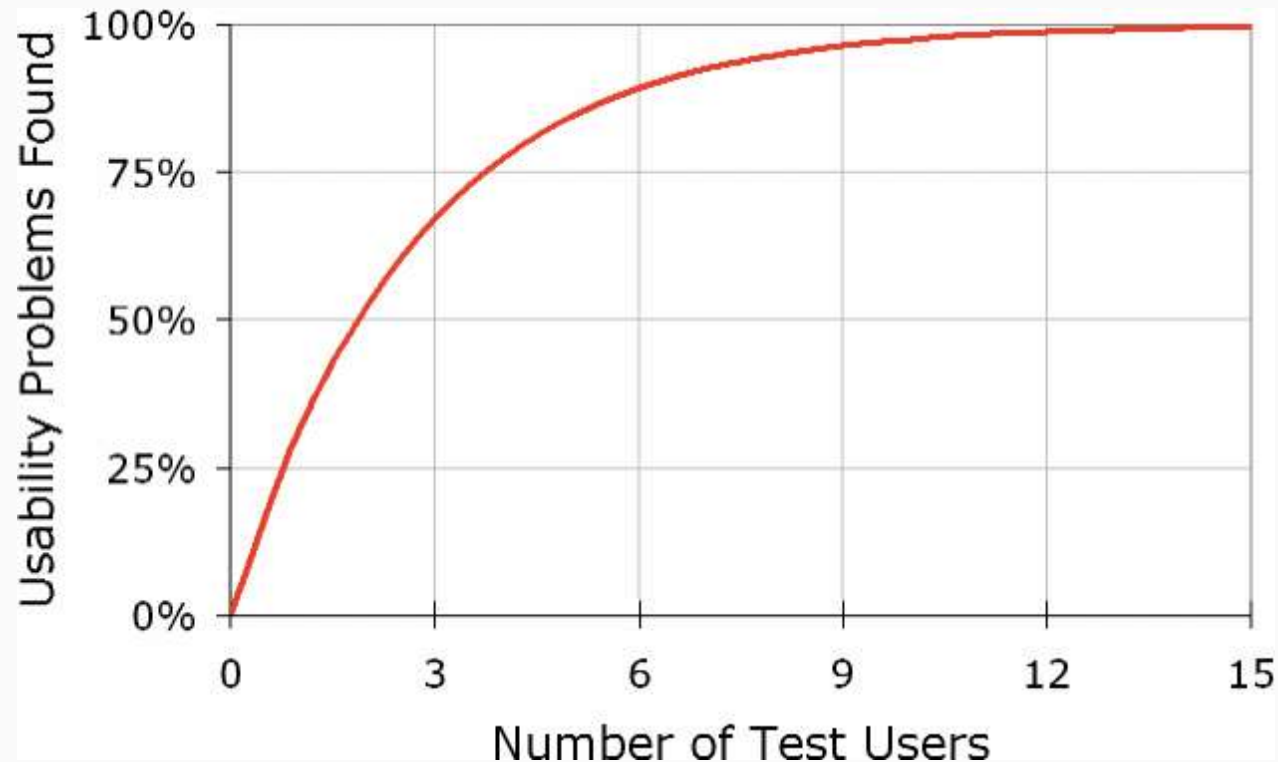
- Three types of data were collected:
 - Satisfaction with Painpad was based on questionnaire responses
 - Patients' compliance with the two-hour routine
 - How data collected from Painpad compared with data collected by nurses
- Data showed:
 - Satisfaction with Painpad 4.63 on Likert scale
 - Patients' compliance was mixed: some liked it while others disliked or didn't notice the prompts
 - Patients recorded more scores with Painpad than through the nurses

How many participants is enough for user testing?

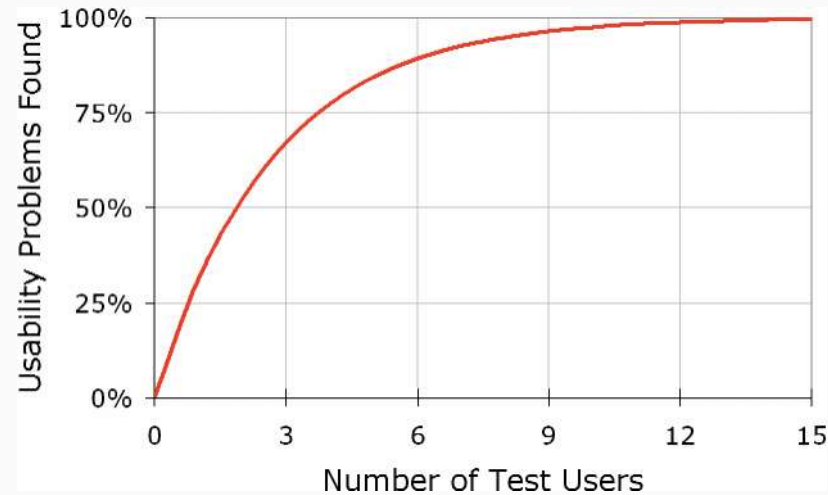
- The number is a practical issue
- Depends on:
 - **Schedule** for testing
 - **Availability** of participants
 - **Cost** of running tests
- Typically **5-10 participants** if possible
- Some experts argue that testing should continue until no new insights are gained
- Others suggest that 5 participants is enough to identify any serious problems

Usability testing: Why Need to Test with 5 Users?

The typical value of L is 31%, averaged across a large number of projects studied. Plotting the curve for $L = 31\%$ gives the following result:



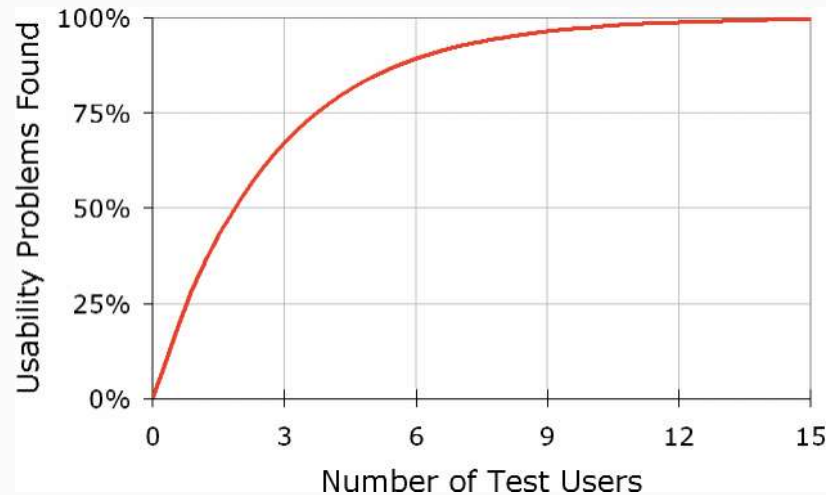
Usability testing: Why Need to Test with 5 Users?



Justify Why 5 users is enough ?

1. As soon as you collect data from **a single test user**, your **insights shoot up** and you have already learned almost a third of all there is to know about the usability of the design.
2. When you test **the second user**, you will discover that this person does some of the same things as the first user, so there is **some overlap in what you learn**. People are definitely different, so there will also be **something new** that the second user does that you did not observe with the first user. So the second user **adds some amount of new insight**, but not nearly as much as the first user did.

Usability testing: Why Need to Test with 5 Users?

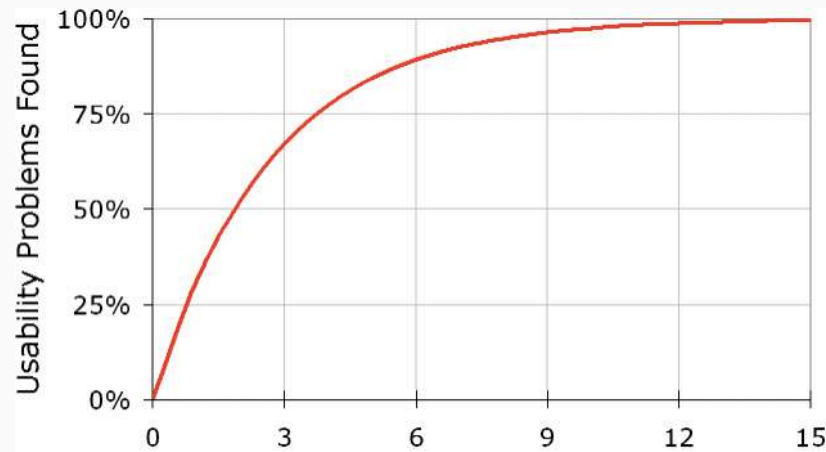


Justify Why 5 users is enough ?

3. The **third user** will **do many things** that you already observed with the first user or with the second user and even some things that you have already seen twice. Plus, of course, the third user will **generate a small amount of new data**, even if not as much as the first and the second user did.

4. As you **add more and more users**, you **learn less and less** because you will **keep seeing the same things again and again**. There is no real need to keep observing the same thing multiple times, and you will be very motivated to go back to the drawing board and redesign the site to eliminate the usability problems.

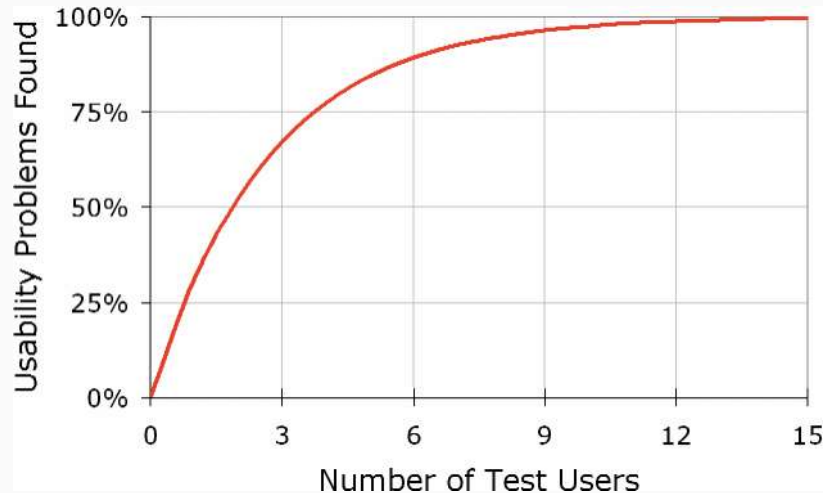
Usability testing: Why Need to Test with 5 Users?



Justify Why 5 users is enough ?

5. After the **fifth user**, you are **wasting your time** by observing the **same findings repeatedly** but not learning much new.

Usability testing: Why Need to Test with 5 Users?



When you need more users for usability testing?

When To Test More Users?

You need to test additional users when a website has several **highly distinct groups of users**. The formula only holds for comparable users who will be using the site in fairly similar ways.

How many participants is **enough** for other types of evaluation?

- **Experiments test hypotheses** to **discover new knowledge** by investigating the **relationship between two or more variables**
- The **number of participants** needed **depends on the type of experiment** being conducted
- It is advisable **to consult a statistician** before deciding
- In-the-wild studies also vary from a few people in a home to a software team to a whole community

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
- Often many participants
- Results validated statistically
- Must be replicable
- Strongly controlled conditions
- Experimental design
- Scientific report to scientific community

Summary

- **Usability testing** takes place in **controlled spaces** - usability labs, temporary labs
- Usability testing focuses on **performance measures** - how long and how many errors are made when completing a set of predefined tasks
- **Indirect observation** (video and keystroke logging), user satisfaction questionnaires, and interviews are also collected
- **Remote testing** has been conducted since the early 1990s but it became important during the Covid19 pandemic
- Remote testing uses **portable equipment** - video and audio recording using smart phones, mobile eye-tracking and automated keystroke logging

Summary (cont.)

- Long-term studies of several weeks are used to evaluate complex products that participants need time to learn and use in their own work
- Experiments test a hypothesis by manipulating certain variables while keeping others constant
- The experimenter controls independent variable(s) in order to measure changes in the dependent variable(s)
- In-the-wild studies are carried out in natural settings to discover how people interact with technology in the real world
- In-the-wild studies involve deployment of prototypes or technologies in natural settings
- Sometimes the findings of in-the-wild studies are unexpected, especially for studies that explore how novel technologies are used by participants in their own homes, places of work, or outside