

INFOMSCIP 2019-2020

lectures 12&13

Oct 14&17, 2019

Major references used for these slides:

- Alan Dix et al., textbook on “Human-Computer Interaction”, chapter 9: “Evaluation techniques”
- Michael Rohs, course on “Mobile Interaction”, TUM, section on user evaluation
- J. Borchers’ course on “Current Topics in Media Computing and Human-Computer Interaction” at RWTH Aachen, <https://hci.rwth-aachen.de/cthci>

Plus various others (referenced on the slides)

Disclaimer: These slides may contain copyrighted material that is used for pure educational purposes based on the Fair Use policy.

http://en.wikipedia.org/wiki/Fair_use

Scientific perspectives on GMT (INFOMSCIP)

User studies & HCI

- Human-computer interaction
Intro, research contributions in HCI, GMT relation, subjective / objective, qualitative / quantitative
- HCI evaluation methods involving experts
Cognitive walkthrough, heuristic evaluation, others
- Methods involving end users: quantitative
Controlled experiments (see last time)
- Methods involving end users: qualitative
Observation (recording, think aloud, ...), querying (interviews, questionnaires, ...), ethical issues (treating participants, ...)

Common research approaches in GMT

Fundamental research

Not related to specific data, nor to users.
Answers to universal questions within well-known/accepted scientific frameworks.

Experimental research

Done on a data set from the real world (by measurements) or generated (synthetic data).
Answers questions about the test data.
Can be a theoretical/fundamental or applied.

User study research

Studies or observes **humans** and their behavior.
Answers questions about users.
Can also be fundamental or applied.

Even for fundamental research, **humans** can be relevant

- Judging outcome (or defining it; researchers are humans, too), e.g., quality, difficulty (puzzle games)
- Measure experience, enjoyment, ... (e.g., in relation to difficulty)

Measured **data** can be **about users and their behavior** (or influenced or created by them).

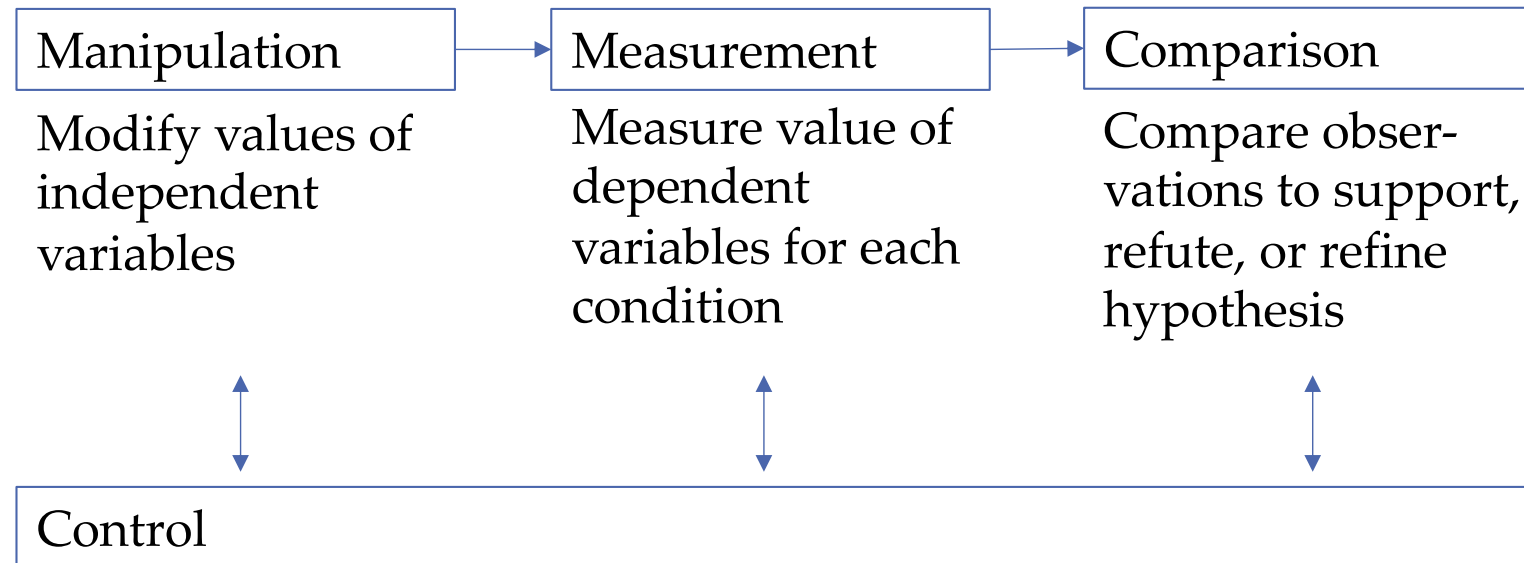
Users are needed to observe and gather data.

Basic elements of an experimental study

*For example:
Input mode
(tilt vs. touch)*

*For example:
game score*

*For example: game
score for tilt vs. touch*



For example: age, gender, experience, location, ...

In computer science (and esp. GMT), we are often interested in **building systems**, e.g., a new type of game or the perfect video search engine.

Keep in mind that we are interested in **universal knowledge** about how to build better systems.

Distinguish pure implementations from scientific research.

⇒ Specify “new”, “perfect” etc. with concrete, measurable characteristics.

Possible criteria:

- Performance
- Usability
- Experience
- ...

This can involve many user studies from natural and social sciences. But mostly from **human-computer interaction**.

Human-computer interaction (HCI) applies user studies to:

- test functional capabilities
- test the impact on users
- identify problems

of a system or interaction design.

From Wikipedia.org

(Human-computer interaction):

As a field of research, **human–computer interaction** is situated at the intersection of computer science, behavioral sciences, design, media studies, and several other fields of study.

HCI studies the **usage of technology**

- Empirical research is often applied in HCI

HCI is also closely related to the **design & development** of technology

- Other research methods are applied as well
- Borders between research and development sometimes not strict

Contribution types in HCI

Empirical research contributions

Artifact contributions

Methodological contributions

Theoretical contributions

Database contributions

Survey contributions

Opinion contributions

Wobbrock, J.O. and Kientz, J.A., 2016.
Research contributions in human-
computer interaction.
interactions, 23(3), pp.38-44.

What does it mean to do research in human computer interaction?

Wobbrock and Keintz classify HCI research into these seven categories of contributions.

They have been adopted by the ACM CHI conference series to categorize their submissions.

If you will ever do an HCI-related project, it is highly recommended that you have a look at this paper.

For each contribution type, it also lists some example papers that can serve as inspiration, help, or even blueprint for your own research.

Contribution types in HCI

Empirical research contributions

Artifact contributions

Methodological contributions

Theoretical contributions

Database contributions

Survey contributions

Opinion contributions

Wobbrock, J.O. and Kientz, J.A., 2016.
Research contributions in human-
computer interaction.
interactions, 23(3), pp.38-44.

EMPIRICAL RESEARCH CONTRIBUTIONS

Empirical research contributions are the backbone of science. They provide new knowledge through findings based on observation and data gathering.

Data may be qualitative or quantitative, aspiringly objective or unapologetically subjective, from the laboratory or from the field.

Empirical research contributions are **evaluated** mainly on the importance of their findings and on the soundness of their methods.

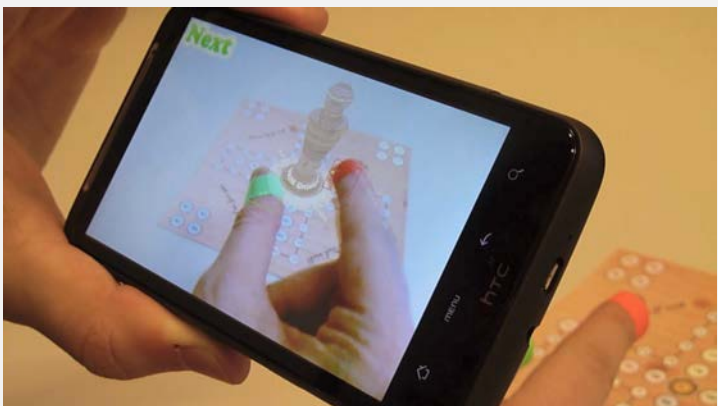
This is a very important contribution type for GMT, too (see various examples in preceding lectures)



Psychophysical
experiment testing
human perception
under varying stimuli



Empirical user study
to verify performance
for varying types of
feedback with
different modalities.



User study verifying
different implementation
options for finger-based
AR interaction.

Also: remember different
“types/levels” of user studies

Study of basic interaction
and perception aspects



Full system
implementation

Contribution types in HCI

Empirical research contributions

Artifact contributions

Methodological contributions

Theoretical contributions

Database contributions

Survey contributions

Opinion contributions

Wobbrock, J.O. and Kientz, J.A., 2016.
Research contributions in human-
computer interaction.
interactions, 23(3), pp.38-44.

ARTIFACT CONTRIBUTIONS

HCI is driven by the creation and realization of interactive artifacts. Whereas empirical contributions arise from descriptive discovery driven activities (science), artifact contributions arise from generative design-driven activities (invention).

Artifact research contributions are **evaluated** according to the type of artifact that gave rise to them. They are often accompanied by empirical studies but do not have to be, and sometimes should not be.

This contribution type might be even more important for GMT than for HCI, since often we introduce new concepts, ideas, ... (e.g., a new interaction method for games)

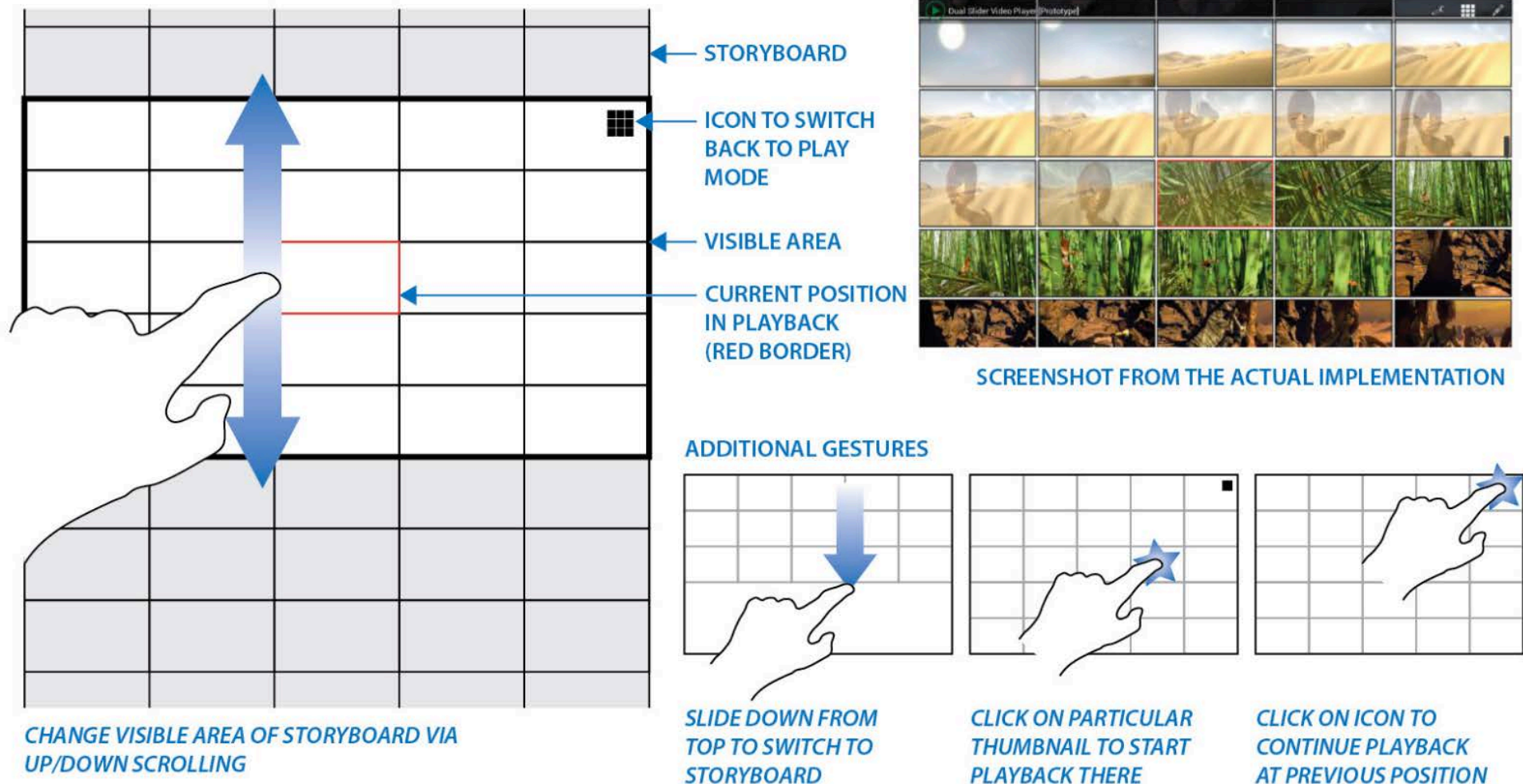
https://www2.projects.science.uu.nl/cs-gmt/index.php?r=project/view&id=52&title_slug=drawing-in-a-virtual-3d-space-introducing-vr-drawing-in-elementary-school-art-education

Wendy Bolier et al. 2018. **Drawing in a Virtual 3D Space - Introducing VR Drawing in Elementary School Art Education.** In Proceedings of the 26th ACM international conference on Multimedia (MM '18).

Monday, October 14, 2019

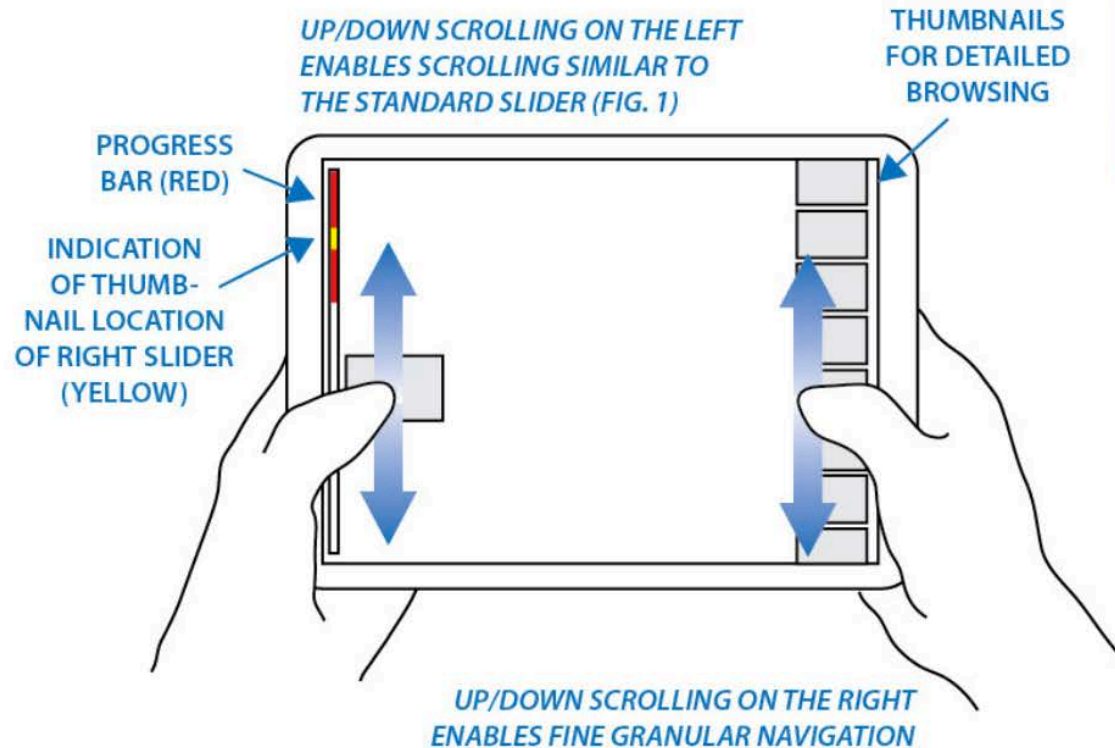
UI design for mobile video – Example

Storyboard-based interaction



Wolfgang Hürst, Miklas Hoet (2015) **Sliders Versus Storyboards – Investigating Interaction Design for Mobile Video Browsing**, Proceedings of MMM 2015

UI design for mobile video – Example

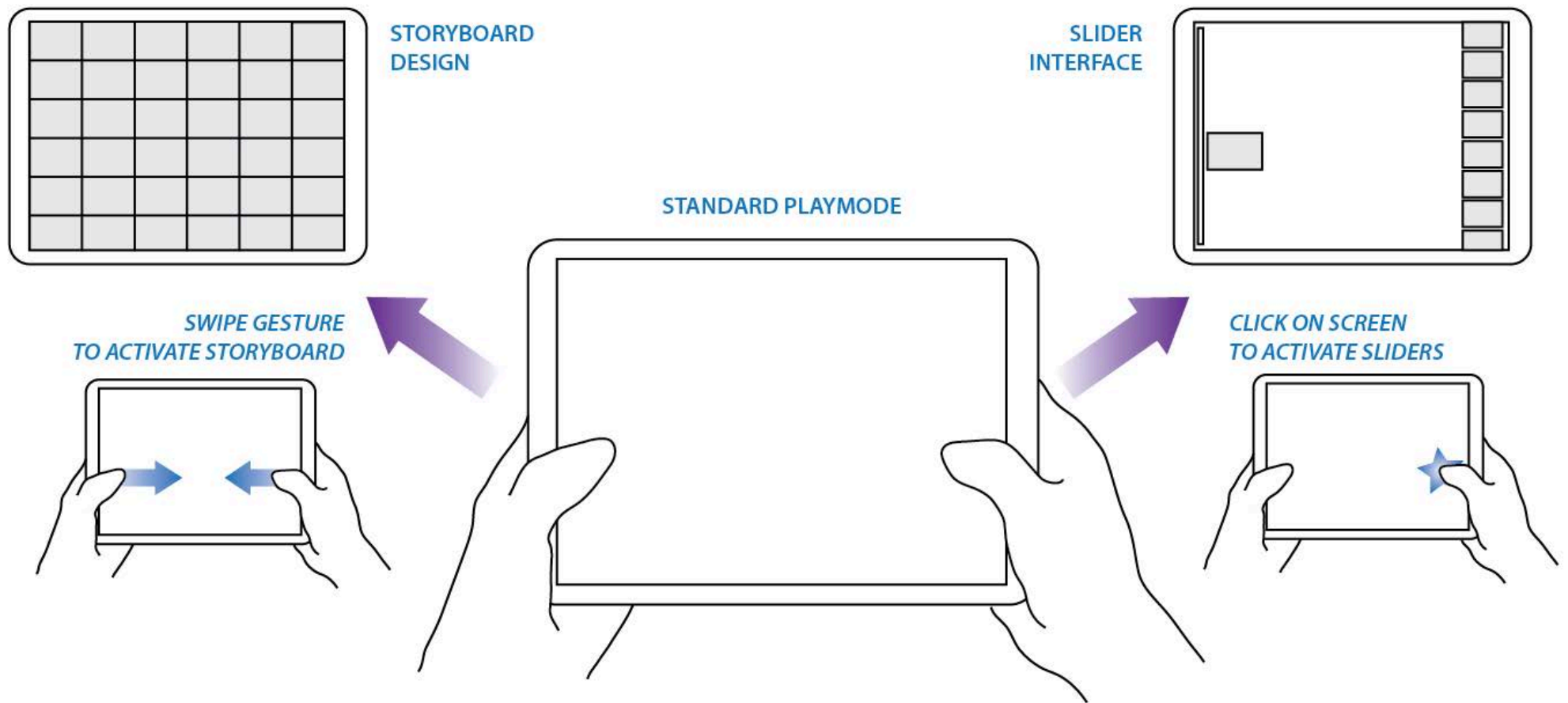


Two-handed slider-based interaction

Wolfgang Hürst, Miklas Hoet (2015) **Sliders Versus Storyboards – Investigating Interaction Design for Mobile Video Browsing**, Proceedings of MMM 2015

UI design for mobile video – Example

Evaluation showed neither a subjective nor objective preference
⇒ Thus combine both approaches into one UI



Wolfgang Hürst, Miklas Hoet (2015) **Sliders Versus Storyboards – Investigating Interaction Design for Mobile Video Browsing**, Proceedings of MMM 2015

Contribution types in HCI

Empirical research contributions

Artifact contributions

Methodological contributions

Theoretical contributions

Database contributions

Survey contributions

Opinion contributions

Wobbrock, J.O. and Kientz, J.A., 2016.
Research contributions in human-
computer interaction.
interactions, 23(3), pp.38-44.

METHODOLOGICAL CONTRIBUTIONS

Methodological research contributions create new knowledge that informs how we carry out our work.

Such contributions may improve research or practice. They may influence how we do science or how we do design.

Methodological research contributions are **evaluated** on the utility, reproducibility, reliability, and validity of the new method or method enhancement.

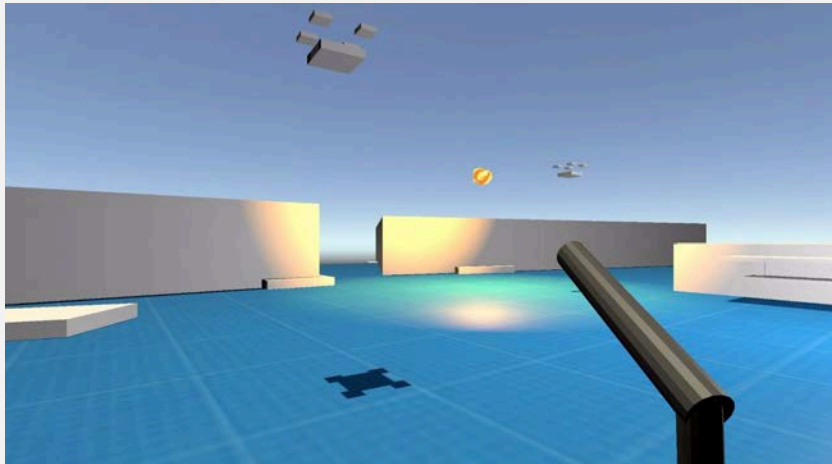
Also in GMT, we find such “research about how to do research”. See, for example, the comparisons of lab versus field studies in relation to mobile interaction from last time. Or the introduction of the “The Game Experience Questionnaire” (GEQ) below.

Context: Presence in VR

Presence, the experience of 'being' or 'acting' when physically situated in another place, is a fundamental characteristic of VR.

Measuring presence is vital for VR research and development.

Typically repeatedly assessed through questionnaires completed after leaving a VR scene.



Assume the following example:

Abstract and realistic scene of a first-person shooter game developed to induce different levels of presence.

(Informal) question:

How do these implementations influence presence? Which one is 'better'?



V. Schwind et al. 2019. **Using Presence Questionnaires in Virtual Reality**. ACM CHI 2019.

Problem: Requiring participants to leave and re-enter the VR after testing each condition to fill out the questionnaire costs time and can cause disorientation.

(Informal) question: Is it necessary, or can't we just do it in VR?



Image: Virtual (l) and real (r) environment with questionnaire and input controller.

Study to investigate the effect of completing presence questionnaires directly in VR

Experiment: 36 participants experienced two immersion levels and completed 3 standardized presence questionnaires in the real world or VR.

Results: No effect on the questionnaires' mean score, but variance significantly depends on realism and if subjects left the VR

Conclusion: Completing questionnaires in VR does not change the measured presence but can increase consistency of the variance.

Contribution types in HCI

Empirical research contributions

Artifact contributions

Methodological contributions

Theoretical contributions

Database contributions

Survey contributions

Opinion contributions

Wobbrock, J.O. and Kientz, J.A., 2016.
Research contributions in human-
computer interaction.
interactions, 23(3), pp.38-44.

THEORETICAL CONTRIBUTIONS

Theoretical research contributions consist of new or improved concepts, definitions, models, principles, or frameworks.

They are vehicles for thought. Whereas methodological contributions inform how we do things, theoretical contributions inform what we do, why we do it, and what we expect from it. Theories may be qualitative or quantitative.

Theoretical research contributions are **evaluated** based on their novelty, soundness, and power to describe, predict, and explain.

Again, such contributions are also common in GMT. See for example the “Foundation of Digital Games” conference series for more theoretical contributions to game research:

<http://www.foundationsofdigitalgames.org/>

Contribution types in HCI

Empirical research contributions

Artifact contributions

Methodological contributions

Theoretical contributions

Database contributions

Survey contributions

Opinion contributions

Wobbrock, J.O. and Kientz, J.A., 2016.
Research contributions in human-
computer interaction.
interactions, 23(3), pp.38-44.

DATASET CONTRIBUTIONS

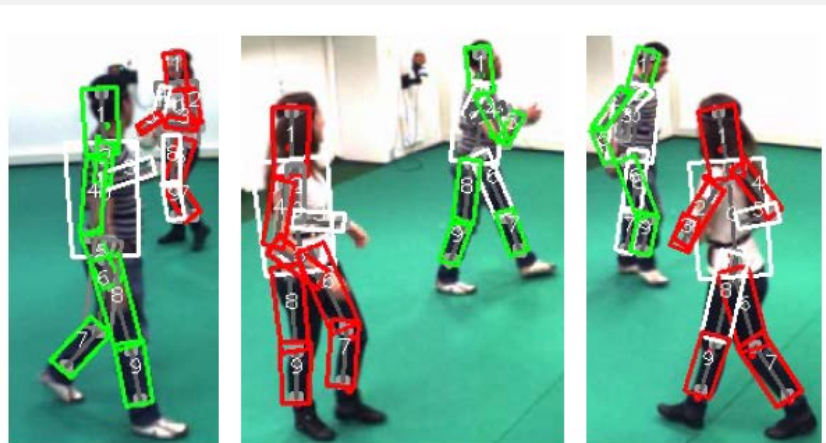
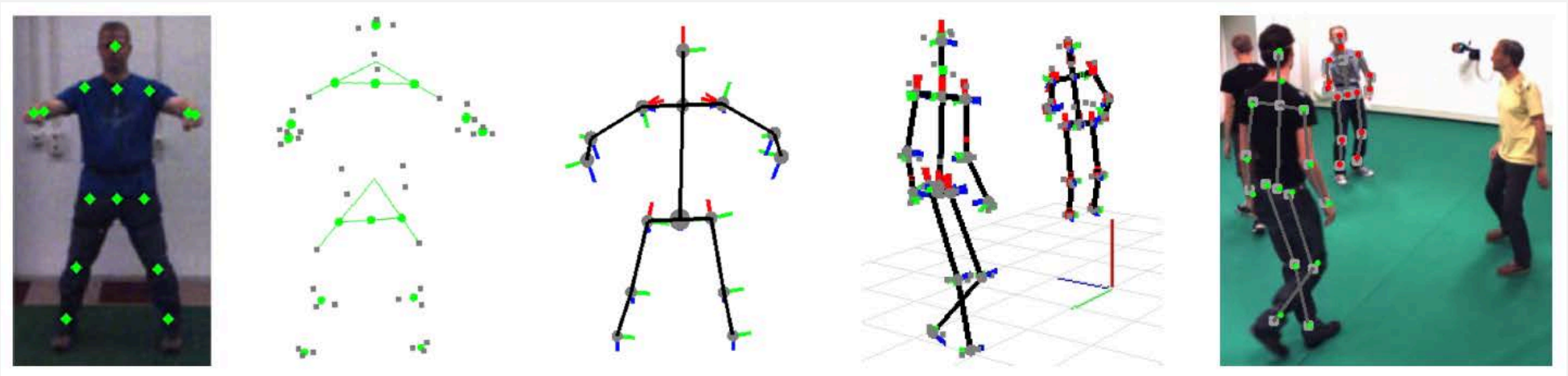
A dataset contribution provides a new and useful corpus, often accompanied by an analysis of its characteristics, for the benefit of the research community.

Benchmark tests may accompany datasets to standardize comparisons.

Dataset research contributions are **judged** favorably by the extent to which they supply the research community with a useful and representative corpus against which to test and measure.

Datasets are also very important for GMT; see TRECVID for an example from media technology research, and the following slide for an example from game research.

Van der Aa, N.P., Luo, X., Giezeman, G.J., Tan, R.T. and Veltkamp, R.C., 2011. **Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction.** In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)* (pp. 1264-1269). IEEE.



Benchmark dataset for multiple people tracking. Contains video recordings as well as synchronized motion capture data (which can be used as ground truth or training data for video camera-based tracking approaches).

Contribution types in HCI

Empirical research contributions

Artifact contributions

Methodological contributions

Theoretical contributions

Database contributions

Survey contributions

Opinion contributions

Wobbrock, J.O. and Kientz, J.A., 2016.
Research contributions in human-
computer interaction.
interactions, 23(3), pp.38-44.

SURVEY CONTRIBUTIONS

Survey research contributions and other meta-analyses review and synthesize work done on a research topic with the goal of exposing trends and gaps. Survey contributions are appropriate after a topic has reached a certain level of maturity.

Survey research contributions, and meta-analyses in general, are **evaluated** based on how well they organize what is currently known about a topic and reveal opportunities for further research.

*Obviously, surveys are also important in the field of GMT. See this one for example:
Thomas B.H.: A survey of visual, mixed, and augmented reality gaming.
Computers in Entertainment (CIE).
2012 Oct 1;10(1):3.*

Contribution types in HCI

Empirical research contributions

Artifact contributions

Methodological contributions

Theoretical contributions

Database contributions

Survey contributions

Opinion contributions

Wobbrock, J.O. and Kientz, J.A., 2016.
Research contributions in human-
computer interaction.
interactions, 23(3), pp.38-44.

OPINION CONTRIBUTIONS

Opinion research contributions, also called essays or arguments, seek to change the minds of readers through persuasion. Although the term opinion might suggest a less-than-scientific effort, in fact, opinion contributions draw upon many of the above contribution types to make their case.

Opinion contributions are considered a separate research contribution type not because they lack a research basis, but because their goal is to persuade, not just inform.

Opinion research contributions are **evaluated** on the strength of their argument.

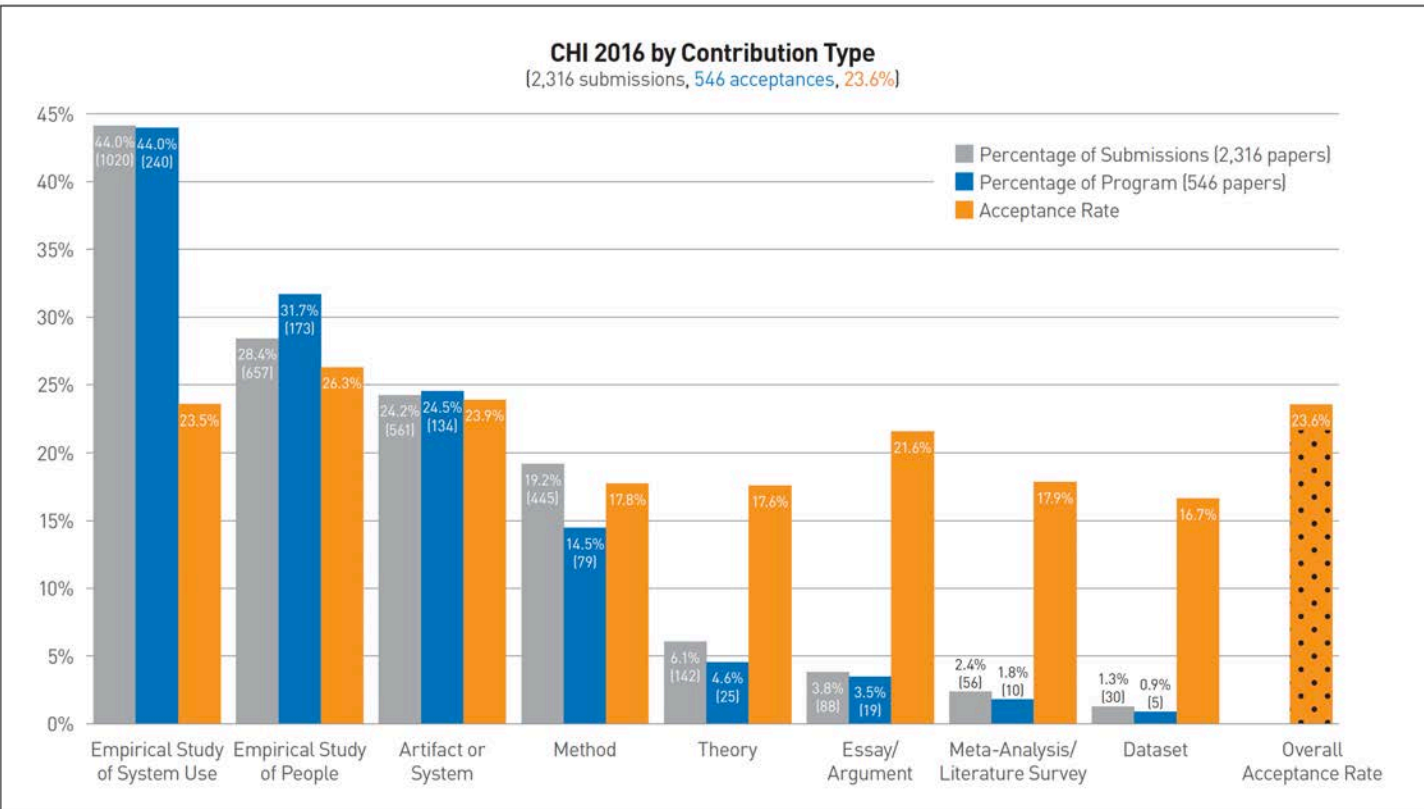
In general, opinion papers are less common but equally important as surveys.

Figure 1. Optional checkboxes for the eight CHI 2016 contribution types. Authors could select none, one, or more than one.

What is the primary contribution type of this paper?

<input type="checkbox"/> Artifact or System	<input type="checkbox"/> Essay/Argument
<input type="checkbox"/> Dataset	<input type="checkbox"/> Meta-Analysis/Literature Survey
<input type="checkbox"/> Empirical study that tells us about how people use a system	<input type="checkbox"/> Method
<input type="checkbox"/> Empirical study that tells us about people	<input type="checkbox"/> Theory

Figure 2. CHI 2016 submissions and acceptances by contribution type, sorted by descending number of submissions.



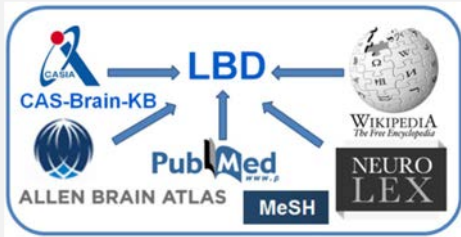
Conclusion: HCI research studies various aspects of human-computer interaction in various ways. But obviously, all of them involve humans (and computers, or better: technology).

User studies in **other domains than HCI** include, but are not limited to:

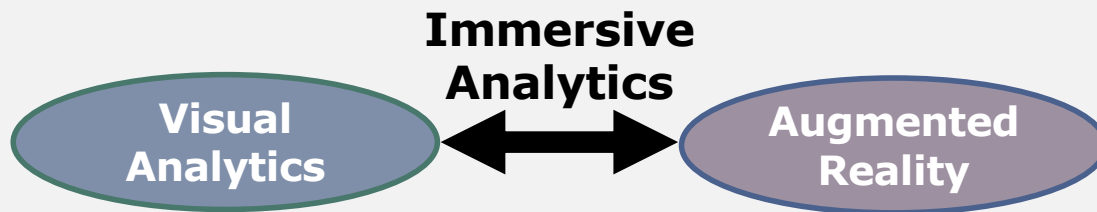
- **Theoretical research**
(e.g., judge difficulty of puzzle games)
- **Computer graphics**
(e.g., qualitative judgements)
- **Visualization**
(e.g., performance tests; see next slide)
- **Virtual reality**
(e.g., perception, immersion, ..., but also performance)
- ...



DatAR: ongoing research project in our group in the field of immersive analytics (visualizing medical data in AR)



Linked Brain Data (LBD): An existing platform for neuroscientists to extract, merge, connect, and analyze large scale brain and neuroscience information from medical publications.



Our goal: use immersive analytics to improve data analysis and relation finding



Mockup / design idea (not implemented yet)

Prototype / idea: A visualization environment DatAR that integrates data into physical spaces using AR

Conclusion: HCI research studies various aspects of human-computer interaction in various ways. But obviously, all of them involve humans (and computers, or better: technology).

User studies in **other domains than HCI** include, but are not limited to:

- **Theoretical research**
(e.g., judge difficulty of puzzle games)
- **Computer graphics**
(e.g., qualitative judgements)
- **Visualization**
(e.g., performance tests)
- **Virtual & mixed reality**
(e.g., perception, immersion, ..., but also performance)
- ...

Notice that these are all GMT-related areas of computer science

Human involvement in HCI-related research methods

Experts

- Cognitive walkthrough
- Heuristic evaluation
- Model-based evaluation
- Literature review

Evaluating the design
(including mockups & prototypes)

Representative samples of potential end users

Qualitative

- Observational techniques
- Query techniques

Quantitative

- Controlled experiments

Evaluating the implementation

Note: exceptions exist and borders are not always strict.

Science must be reproducible (thus: measures). This is usually easier when testing performance, but hard for subjective matters (e.g., experience, usability, ...).

Subjective

- Methods rely on the interpretation of the subject's input
- May provide information that cannot be gathered with objective methods
- Often rely on knowledge and expertise of participant (e.g., experts in cognitive walkthroughs and heuristic evaluations)
- Evaluator bias can be a problem (thus, e.g., use multiple experts)

Objective

- Interpretation of data is independent of subject
- Should produce repeatable results, independent of participants
- May not identify all issues and can lack detailed feedback (e.g., on user experience)

Often, a **combination** is used, e.g., a controlled experiment that measures data objectively and complements it with subjective feedback (e.g., via subsequent interviews)

Quantitative and objective are related aspects, but not the same. The same goes for qualitative and subjective.

	Quantitative	Qualitative
Objective	<p>“The chip of my computer is 2 GHz.”</p> <p>“It took 30 sec to solve the task with this approach.”</p>	<p>“Yes, I own a computer.”</p> <p>“Yes, I solved the task with this approach.”</p>
Subjective	<p>“On a scale from 1-10, my computer scores 7 in terms of its ease of use.”</p> <p>“In terms of speed, I would rate this approach as 7 on a scale from 1-10.”</p>	<p>“I think computers are too expensive.”</p> <p>“The approach allowed me to solve the task quite fast.”</p>

From <https://www.userfocus.co.uk/articles/datathink.html> (blue parts have been added)

Many HCI methods strive to evaluate subjective matters (usefulness, usability, experience, ...) with objective methods (measures!)

Quantitative: values that can be measured & expressed numerically

- Different types exist (discrete/continuous, ...)
- Usually analyzed with statistical means (averages, distributions, significance testing)

Qualitative, e.g., comments & observations

- Can be subjective & biased
- Not directly quantifiable, but means to measure exist, too, e.g., categorize and encode for gaining general results

Separation not always strict (e.g., quantitative description of qualitative data via questionnaires; usage of operational definitions).

It can make sense to gather both (e.g., qualitative data to gather insight into quantitative results, comparison of perceived versus actual performance, etc.).

Operational definitions

Goal: Describe a 'fuzzy' characteristic via quantifiable and measurable parameters.

Examples:

A "healthy baby"

⇒ Body weight / size at birth

A "good driver"

⇒ Years driven without accident

"Improved situational awareness of drivers"

⇒ fewer missed targets (braking car in front, people crossing street) in a driving simulator task

"Better user friendliness of an interface"

⇒ higher satisfaction rating on questionnaire, shorter task completion time compared to state-of-the-art interface

Phrasing often found in publications:

In the present study, we define situational awareness of drivers as XY.

We measure game play experience via the standardized questionnaire X by [Y], which ...

Human involvement in HCI-related research methods

Experts

- Cognitive walkthrough
 - Established evaluation approaches in HCI
 - Usually done by experts
 - Usually focused on the design
 - Usually focused on qualitative feedback
- Heuristic evaluation
- Model-based evaluation
- Literature review

Evaluating the design
(including mockups &
prototypes)

Experts

- Cognitive walkthrough
- Heuristic evaluation
- **Model-based evaluation**
- Literature review

Evaluating the design
(including mockups &
prototypes)

Model-based evaluation

Several theoretical models exist that offer a framework for design and evaluation

Examples

- GOMS (goals, operators, methods, selection rules)
- KLM (keystroke-level model)
- Design rational
- Design patterns

Experts

- Cognitive walkthrough
- Heuristic evaluation
- Model-based evaluation
- **Literature review**

Evaluating the design
(including mockups &
prototypes)

Literature review / review-based evaluation

Search literature for evidence
for or against aspects of your design

Advantage

- Saves own experiments

Potential problem

- Results only carry over reliably
if context is very similar

Might not be considered “real” science,
but rather engineering or development.

Note though that this is often implicitly
part of experiment design in empirical
research (e.g., when setting parameters)

Cognitive walkthrough

Analytical method for early design and existing systems.

Basic idea: let expert do detailed review of a sequence of actions.

Evaluator is not an end user, but an **expert** (designer or cognitive psychologist; technique originated from psychology, now adapted for designers).

Goal is to judge learnability and ease of use.

Does the system help users to get from goals to intentions and actions?

Procedure: Step through each action and ask ...

- Is the effect of the action the same as the user's goal at that point?
- Will users see that the action is available?
- Once users find the action, will they know it is the right one?
- After the action is taken, will users understand the feedback?

(Simulation of "learning through exploration")

Cognitive walkthrough

Four things are needed:

1. Description of the prototype or system (or interface).
2. Description of the task the user is to perform with the system.
Example: Program the VCR to time-record a program starting at 18:00 and finishing at 19:15 on channel 4 on October 10, 2018
3. List of interface actions to complete the task with the system.
4. User profile, i.e., identification of who the users are, their experience and expected knowledge.

Doing the actual walkthrough:

- Analyze process of performing the actions using above questions

Written questions capture psychological knowledge and guide the tester

Heuristic evaluation

1. Choose usability heuristics, collection of usability principles.
E.g., **Nielsen's ten usability guidelines**
2. Go through each task and check whether guidelines are followed.
3. Severity rating for each problem (Nielsen).
 - 0 = I don't agree this is a problem at all
 - 1 = cosmetic problem
 - 2 = minor usability problem, low priority to fix
 - 3 = major usability problem, high priority to fix
 - 4 = usability catastrophe, imperative to fix before release

Advantage:

- Quick and cheap

Disadvantage:

- Subjective

Why You Only Need to Test with 5 Users, Jakob Nielsen, 2000

Summary: Elaborate usability tests are a waste of resources. The best results come from testing no more than 5 users and running as many small tests as you can afford.

Note that this statement is often misunderstood and falsely interpreted!

<https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/>

Heuristic evaluation: Jakob Nielsen's 10 usability principles

Note that different versions of this exist. This is the latest one.

<https://www.nngroup.com/articles/ten-usability-heuristics/>

Visibility of system status

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.

Match between system and the real world

The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

User control and freedom

Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support undo and redo.

Consistency and standards

Users should not have to wonder whether different words, situations, or actions mean the same thing.

Error prevention

Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.

Recognition rather than recall

Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.

Flexibility and efficiency of use

Accelerators — unseen by the novice user — may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.

Aesthetic and minimalist design

Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.

Help users recognize, diagnose, and recover from errors

Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.

Help and documentation

Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Human involvement in HCI-related research methods

Experts

- Cognitive walkthrough
- Heuristic evaluation
- Model-based evaluation
- Literature review

Evaluating the design
(including mockups &
prototypes)

Note: plenty of other guidelines
and standards exist that could
be used for expert evaluations.

Human involvement in HCI-related research methods

Experts

- Cognitive walkthrough
- Heuristic evaluation
- Model-based evaluation
- Literature review

Evaluating the design
(including mockups & prototypes)

Note: plenty of other guidelines and standards exist that could be used for expert evaluations.

Representative samples of potential end users

Qualitative

- Observational techniques
- Query techniques

Quantitative

- Controlled experiments

Evaluating the implementation

Note: exceptions exist and borders are not always strict.

Evaluations with end users: key issues

Setting goals

Decide how to analyze data once collected.
This should follow from your research question.
Yet, make sure to carefully plan it *before* gathering the data.

Relationship with participants

- Clear and professional.
- Protect privacy
- Informed consent form when appropriate
(signed agreement between evaluator and participant)

Triangulation

- Use more than one approach
- Use different perspectives to understand a problem or situation

Iterate

- If questions reveal that goal was not sufficiently defined:
refine it, repeat tests

Dealing with participants

Tests are often uncomfortable for the one tested.
(Pressure to perform, mistakes, competitive thinking)

Make sure that they are aware that they are not tested
but the system is (by them).

Anonymity is common and most often makes sense.

Inform them that they can stop at any time without providing any reasons.

Providing this information in written form is often useful (including signed consent, e.g., about recording or otherwise collecting data).

Treat them with respect at all times - Before, during, and after the test.

Note: this sounds obvious and easy, but it sometimes isn't.

E.g., a company hires you to study their work process to improve it and save money. Employees participating in the study are guaranteed anonymity. You realize that one of them is bad at his job and his mistakes cost the company lots of money. Is your loyalty with the company who pays you to save them money, or with the employee who you promised anonymity and who might get fired if you show your results to the company?

Qualitative approaches

(In HCI often combined with quantitative ones)

Generally more concerned with reasons underlying human behavior

Generally focused more on human and subjective aspects
(e.g., feeling, liking, sentiment, ...)

Number of subjects often lower
(if purely qualitative; be careful if both, e.g.,
when including quantitative measures of qualitative statements,
for example via questionnaires).

Various ways to gather it, e.g.,

- Observations, including think aloud techniques, interviews
- Empirical measures (questionnaires, ...,
but also sensors, e.g., heart rate as indication for stress)

How to gather qualitative, observational data

Recording of observations, interactions, comments, etc.

- Handwritten notes (from observer)
 - Pros: cheap and easy, non-intrusive
 - Cons: easy to miss details, writing is slow
- Notes from the participant (e.g., diary for long term studies)
- Audio or video recordings
 - Video: film user and screen (2 cameras if needed)
 - Pros: detailed recording for later analysis
 - Cons: intrusive, users might feel uncomfortable and not act natural
 - Make sure to get user consent!
- Computer logs
 - Screen captures (detailed, but might be hard to analyze, lots of work).
 - Interaction logs (hint: log as much as you can, think about format that is easy to process)

In practice: often a mixture of approaches is used (e.g., audio recording for think aloud technique plus computer logs to recreate interface status).

How to gather qualitative comments from the subject (1/2)

Silent observation

- Observer watches user doing the experiment in the lab or elsewhere
- No communication takes place during the tests

Pros: Helps discover big problems, no influence of participant

Cons: No understanding of decision process, user's mental models, opinions, or feelings

Think aloud technique (most common method in industry)

- Observer is silent, but user is asked to say aloud:
 - What he/she thinks is happening (state)
 - What he/she is trying to achieve (goals)
 - Why he/she is doing something specific (actions)

Pros: Good to get some insight into user's thinking

Cons: Talking is hard while focusing on task.

For some (most?) people it feels weird talking aloud.

Conscious talking can change behavior.

How to gather qualitative comments from the subject (2/2)

Constructive interaction

- Two people work on a task together

Pros: Normal conversation is observed.

More comfortable than think aloud.

Cons: One user's comments or actions can influence the other one's.

Variation of this: different types of partners

- Semi-expert as “trainer”, newbie as “student”
- Student uses UI and asks, trainer answers

Pros: Gives insight into mental models of beginner and advanced user at the same time.

Retrospective testing / post-task walkthroughs

- Observer and participant look at recorded data after the test, user comments on his/her actions retrospectively

Good when talking during the test should be avoided (e.g., when measuring performance).

Often results in concrete suggestions for improvement.

Human involvement in HCI-related research methods

Representative samples of potential end users

Qualitative

- Observational techniques
- Query techniques
 - Interviews
 - Questionnaires

Quantitative

- Controlled experiments

Interviews

Direct and structured way to gather information.

Users can be probed more deeply on interesting issues that arise.

Usually a top-down approach.

General questions first,
more leading questions later (Why ...? What if ...?)

Advantages

- Good at providing high level info (preferences, impressions, attitude).
- May reveal problems not noticed during observation.
- Can help to clarify events if used together with observations.

Questionnaires

Less flexible than interviews, likely less probing.

Can reach wider group of subjects,
take less time to administer,
can be analyzed more rigorously.

Need to be well designed to gather meaningful information.

Types of questions:

- General (e.g., demographics, experience, ...)
- Open ended (e.g., "Can you suggest any improvements to the interface?")
- Scalar (e.g., Likert scale)

It is easy to recover from mistakes.						
Disagree	1	2	3	4	5	Agree

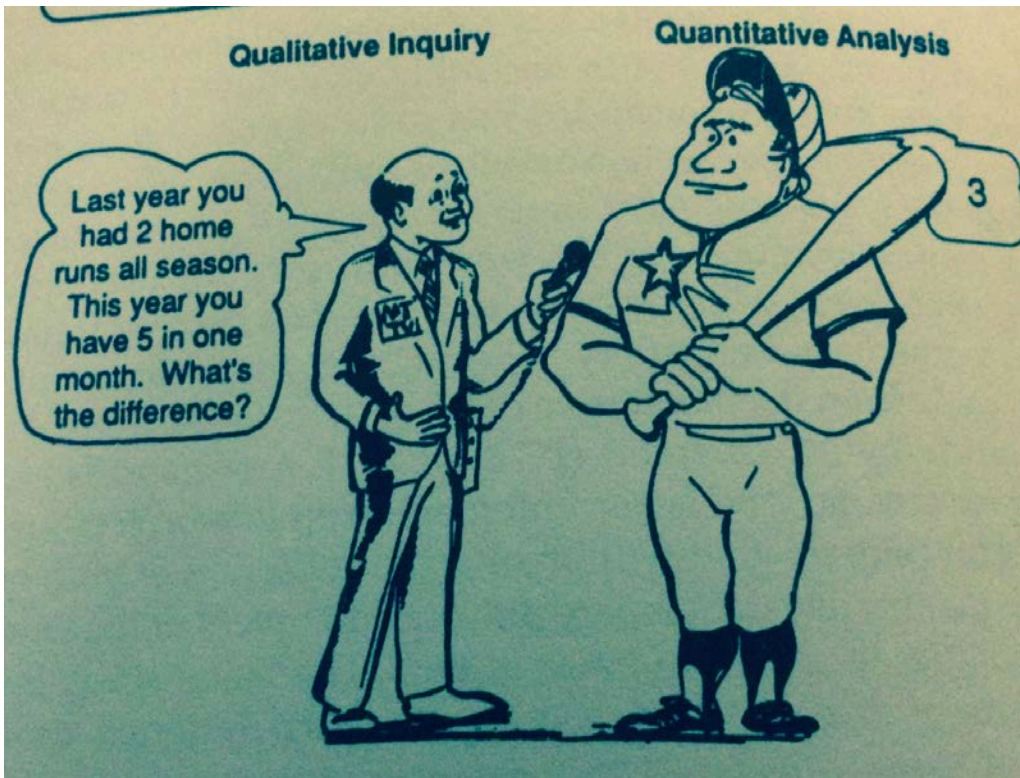
See also, e.g., https://www.cdc.gov/dhds/pubs/docs/CB_February_14_2012.pdf

- Multi-choice (with one or multiple possible options to chose)
- Ranked (e.g., rank preference of tested conditions)

Questionnaires

What to ask and how?

Guidelines exist for the latter.



Often: **goals** are specified via your research question or hypothesis

Questionnaires are designed to answer this question.

Operational definitions are often helpful here.

But questionnaire design and correct phrasing is not easy.

Luckily, many standards exist that can either be used or adapted for your research.

Examples here:

- Usability
- Workload
- User engagement
- Game experience

Usability in HCI (definition)

Learnability

How easy is it for users to accomplish basic tasks the first time they encounter the design?

Memorability

When users return to the design after a period of not using it, how easily can they re-establish proficiency?

Efficiency

Once users have learned the design, how quickly can they perform tasks?

Quantitative entity

Errors

How many errors do users make, how severe are these errors, and how easily can they recover from the errors?

Quantitative entity

Satisfaction

How pleasant is it to use the design?

Qualitative characteristic

Typical Measures of Effectiveness

- Binary task completion
- Accuracy
 - Error rates
 - Spatial accuracy
 - Precision
- Recall
- Completeness
- Quality of outcome
 - Understanding
 - Experts' assessment
 - Users' assessment

Kasper Hornbæk: Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Human-Computer Studies* 64 (2006) 79–102.

Typical Measures of Efficiency

- Time
 - Task completion time
 - Time in mode (e.g., time in help)
 - Time until event (e.g., time to react to warning)
- Input rate (e.g., words per minute, WPM)
- Mental effort (NASA Task Load Index)
- Usage patterns
 - Use frequency (e.g., number of button clicks)
 - Information accessed (e.g., number of Web pages visited)
 - Deviation from optimal solution (e.g. path length)
- Learning (e.g., shorter task time over sessions)

Kasper Hornbæk: Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Human-Computer Studies* 64 (2006) 79–102.

Typical Measures of Satisfaction

- Standard questionnaires (e.g., QUIS)
- Preference
 - Rate or rank interfaces
 - Behavior in interaction (e.g., observe what users choose)
- Satisfaction with the interface
 - Ease-of-use (e.g. 5-/7-point Likert scale: “X was easy to use”)
 - Satisfaction with specific features
 - Before use (e.g., “I will be able to quickly find pages”)
 - During use (e.g., heart period variability, reflex responses)
- Attitudes and perceptions
 - Attitudes towards others (e.g., “I felt connected to X when using...”)
 - Perception of outcome / interaction

Kasper Hornbæk: Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Human-Computer Studies* 64 (2006) 79–102.

Typical Measures of Specific Attitudes

- Annoyance
- Anxiety
- Complexity
- Control
- Engagement
- Flexibility
- Fun
- Liking
- Want to use again

Kasper Hornbæk: Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Human-Computer Studies* 64 (2006) 79–102.

Usability example: System Usability Scale (SUS)

- Developed by DEC Cooperation
- Ten 5-point Likert scales (from “strongly agree” to “strongly disagree”)
- Can be combined in single score (0-100)

1. I think that I would like to use this system frequently

Strongly disagree Strongly agree

1	2	3	4	5

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.

<https://usabilitygeek.com/how-to-use-the-system-usability-scale-sus-to-evaluate-the-usability-of-your-website/>

User satisfaction example: Questionnaire for User Interaction Satisfaction (QUIS)

- Developed by the University of Maryland
- Semantic differential scales
- Components:
 1. Demographics
 2. Overall reaction ratings (6 scales)
 3. Specific interface factors: screen, terminology and feedback, learning, system capabilities
 4. Optional sections
- Long and short forms exist

<http://garyperلمان.com/quest/quest.cgi?form=QUIS>

Questionnaire for User Interface Satisfaction

Based on: Chin, J.P., Diehl, V.A., Norman, K.L. (1988) *Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface. ACM CHI'88 Proceedings*, 213-218. ©1988 ACM. [Abstract] Copying without fee is permitted provided that the copies are not made or distributed for direct commercial advantage, and credit to the source is given. ©1986-1998 University of Maryland. This display is for educational uses only. Commercial use requires a license from the Office of Technology Commercialization: (301) 403-2711 otc@umail.umd.edu. [QUIS Home Page] | [About quest.cgi](#)

Please rate your satisfaction with the system.

- Try to respond to all the items.
- For items that are not applicable, use: NA
- Make sure these fields are filled in: **System:** **Email to:**
- Add a comment about an item by clicking on its icon, or add comment fields for all items by clicking on **Comment All**
- To mail in your results, click on: **Mail Data**

System: Email to:
Optionally provide comments and your email address in the box.

OVERALL REACTION TO THE SOFTWARE		0	1	2	3	4	5	6	7	8	9	NA
1. <input type="checkbox"/>	terrible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	wonderful <input type="radio"/>
2. <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy <input type="radio"/>
3. <input type="checkbox"/>	frustrating	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	satisfying <input type="radio"/>
4. <input type="checkbox"/>	inadequate power	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	adequate power <input type="radio"/>
5. <input type="checkbox"/>	dull	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	stimulating <input type="radio"/>
6. <input type="checkbox"/>	rigid	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	flexible <input type="radio"/>
SCREEN		0	1	2	3	4	5	6	7	8	9	NA
7. Reading characters on the screen <input type="checkbox"/>	hard	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy <input type="radio"/>
8. Highlighting simplifies task <input type="checkbox"/>	not at all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	very much <input type="radio"/>
9. Organization of information <input type="checkbox"/>	confusing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	very clear <input type="radio"/>
10. Sequence of screens <input type="checkbox"/>	confusing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	very clear <input type="radio"/>
TERMINOLOGY AND SYSTEM INFORMATION		0	1	2	3	4	5	6	7	8	9	NA
11. Use of terms throughout system <input type="checkbox"/>	inconsistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	consistent <input type="radio"/>
12. Terminology related to task <input type="checkbox"/>	never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	always <input type="radio"/>
13. Position of messages on screen <input type="checkbox"/>	inconsistent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	consistent <input type="radio"/>
14. Prompts for input <input type="checkbox"/>	confusing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	clear <input type="radio"/>
15. Computer informs about its progress <input type="checkbox"/>	never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	always <input type="radio"/>
16. Error messages <input type="checkbox"/>	unhelpful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	helpful <input type="radio"/>
LEARNING		0	1	2	3	4	5	6	7	8	9	NA
17. Learning to operate the system <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy <input type="radio"/>
18. Exploring new features by trial and error <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy <input type="radio"/>
19. Remembering names and use of commands <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy <input type="radio"/>
20. Performing tasks is straightforward <input type="checkbox"/>	never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	always <input type="radio"/>
21. Help messages on the screen <input type="checkbox"/>	unhelpful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	helpful <input type="radio"/>
22. Supplemental reference materials <input type="checkbox"/>	confusing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	clear <input type="radio"/>
SYSTEM CAPABILITIES		0	1	2	3	4	5	6	7	8	9	NA
23. System speed <input type="checkbox"/>	too slow	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	fast enough <input type="radio"/>
24. System reliability <input type="checkbox"/>	unreliable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	reliable <input type="radio"/>
25. System tends to be <input type="checkbox"/>	noisy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	quiet <input type="radio"/>
26. Correcting your mistakes <input type="checkbox"/>	difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	easy <input type="radio"/>
27. Designed for all levels of users <input type="checkbox"/>	never	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	always <input type="radio"/>
		0	1	2	3	4	5	6	7	8	9	NA

List the most **negative** aspect(s):

1.

2.

3.

List the most **positive** aspect(s):

1.

2.

3.



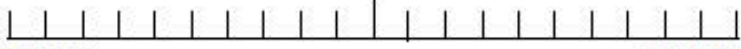



Workload example: NASA Task Load Index (NASA-TLX)

- Developed by NASA
- Subjective, multidimensional assessment tool for rating perceived workload
- Assesses a task, system's or team's effectiveness or other aspects of performance
- Six questions on the following aspects:
 1. Mental Demand
 2. Physical Demand
 3. Temporal Demand
 4. Performance
 5. Effort
 6. Frustration
- Used in a variety of domains (including mobile HCI)

<https://en.wikipedia.org/wiki/NASA-TLX>

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date
<hr/>		
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

User engagement: How to measure?

User engagement is the emotional, cognitive, and behavioral connection that exists, at any point in time and possibly over time, between a user and a resource.

From Attfield et al., 2011: *Towards a science of user engagement (position paper)*

How to measure?

Self-report:

What: Happy, sad, enjoyment, ...
Means: Questionnaire, interview, think-aloud and think after protocols, ...
Attributes: Subjective, short- and long-term, lab & field, small scale

Physiology:

What: Gaze, body heat, mouse movement, ...
Means: EEG, SCL, fMRI, eye tracking, mouse-tracking, ...
Attributes: Objective, short-term, lab & field, small & large scale

Analytics:

What: Click, upload, read, comment, share, ...
Means: Intra and inter-session metrics, data science, ...
Attributes: Objective, short- and long-term, field, large scale

User engagement: Qualitative measures

Self-report:

What: Happy, sad, enjoyment, ...

Means: Questionnaire, interview, think-aloud and think after protocols, ...

Attributes: Subjective, short- and long-term, lab & field, small scale

⇒ *Qualitative measures needed*

Established questionnaires often used in this context include:

- Focused attention questionnaire [O'Brien & Toms, 2010]
- PANAS questionnaire [Watson, Clark & Tellegen, 1988]

User engagement: Focused attention questionnaire

by O'Brien & Toms, 2010

1. I lost myself in this news tasks experience
2. I was so involved in my news tasks that I lost track of time
3. I blocked things out around me when I was completing the news tasks
4. When I was performing these news tasks, I lost track of the world around me
5. The time I spent performing these news tasks just slipped away
6. I was absorbed in my news tasks
7. During the news tasks experince I let myself go

Ratings on a 5-point Likert-scale
from “strong disagree” to “strong agree”

User engagement: PANAS questionnaire by Watson, Clark, and Tellegen, 1988

Answers are used to calculate a **mean positive affective score** and a **mean negative affective score** (by summing up ratings for positive and negative items).

The original paper verified the **reliability** and **validity** of this approach.

It provided a **ground truth**, i.e. for a “normal population”, the mean positive affective score should be 29.7 (SD = 7.9) and the negative affective score should be 14.8 (SD = 5.4)

“You feel this way right now, that is, at the present moment?”

Ratings on a 5-point Likert-scale with

1 = *very slightly or not at all*

2 = *a little*

3 = *moderate*

4 = *quite a bit*

5 = *extremely*

for the following 10 positive and 10 negative items (presented in randomized order):

Distressed, upset, guilty, scared, hostile, irritable, ashamed, nervous, jittery, afraid

Interested, excited, strong, enthusiastic, proud, alert, inspired, determined, attentive, active

Game experience: The Game Experience Questionnaire (GEQ)

https://pure.tue.nl/ws/files/21666907/Game_Experience_Questionnaire_English.pdf



The Game Experience Questionnaire

IJsselsteijn, W.A.; de Kort, Y.A.W.; Poels, K.

Published: 01/01/2013

Document Version

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)