

Fall 2019

CS6501: Topics in Human-Computer Interaction

[http://seongkookheo.com/cs6501\\_fall2019](http://seongkookheo.com/cs6501_fall2019)

# Lecture 8: Qualitative Evaluation 1

Seongkook Heo

September 19, 2019

# What You Previously Learned

- Quantitative Evaluation Methods
- Process of conducting an experiment
  - Research Hypothesis
- Null Hypothesis Significance Testing
  - Statistical Analysis

# Qualitative Research Methods

- Query and Critique
  - Ask people to provide judgements or feedback on a design
  - Excellent for getting initial and continued feedback on a design
  - May not identify all problems that will actually occur
  - Some methods are time consuming
- Observe Users
  - Observe users actually using the system or prototype
  - Typically quick and easy to do
  - Some methods can reveal what a person is thinking



<https://www.youtube.com/watch?v=ELYVpikRNEE>

# Bias

- Qualitative Evaluation is more prone to bias
  - Confirmation bias
  - Unanswerable question bias
  - Social acceptance bias
  - Moderator bias

# Bias

- Qualitative Evaluation is more prone to bias
  - Confirmation bias
    - Avoid sending signal what you'd like the response to be, or with prior judgement, or your own evaluation.  
e.g., many students prefer this book. What do you think?
  - Unanswerable question bias
  - Social acceptance bias
  - Moderator bias

# Bias

- Qualitative Evaluation is more prone to bias
  - Confirmation bias
  - Unanswerable question bias
    - Avoid asking about information that can't be answered reliably.  
e.g., that are highly specific, requiring foresight, or others' experiences.
  - Social acceptance bias
  - Moderator bias

# Bias

- Qualitative Evaluation is more prone to bias
  - Confirmation bias
  - Unanswerable question bias
  - Social acceptance bias
    - Be careful about socially desirable/undesirable questions.  
e.g., requiring answers that people may be reluctant to be honest
- Moderator bias



# Bias

- Qualitative Evaluation is more prone to bias
  - Confirmation bias
  - Unanswerable question bias
  - Social acceptance bias
  - Moderator bias
    - Avoid being too familiar with the participant.  
They may want to please you rather than being honest.

# Bias

- Something to read:

- Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies.  
"Yours is better!": participant response bias in HCI.  
*CHI '12.*

Session: The Tools of the Trade

CHI 2012, May 5–10, 2012, Austin, Texas, USA

## **"Yours is Better!" Participant Response Bias in HCI**

**Nicola Dell<sup>†</sup>**

**Vidya Vaidyanathan<sup>‡</sup>**

**Indrani Medhi<sup>§</sup>**

**Edward Cutrell<sup>§</sup>**

**William Thies<sup>§</sup>**

<sup>†</sup>University of Washington  
nixdell@cs.washington.edu

<sup>‡</sup>San Jose State University  
vidya.vn@gmail.com

<sup>§</sup>Microsoft Research India  
{indranim,cutrell,thies}@microsoft.com

### **ABSTRACT**

Although HCI researchers and practitioners frequently work with groups of people that differ significantly from themselves, little attention has been paid to the effects these differences have on the evaluation of HCI systems. Via 450 interviews in Bangalore, India, we measure participant response bias due to interviewer demand characteristics and the role of social and demographic factors in influencing that bias. We find that respondents are about 2.5x more likely to prefer a technological artifact they believe to be developed by the interviewer, even when the alternative is identical. When the interviewer is a foreign researcher requiring a translator, the bias towards the interviewer's artifact increases to 5x. In fact, the interviewer's artifact is preferred even when it is degraded to be obviously inferior to the alternative. We conclude that participant response bias should receive more attention within the CHI community, especially when designing for underprivileged populations.

### **Author Keywords**

HCI4D; ICTD; demand characteristics; interviewer effects;



Figure 1. Interviewing an auto rickshaw driver in Bangalore, India. When shown two technologies by a foreign interviewer (with translator), rickshaw drivers preferred the one they believed to be developed by the interviewer, even when it was obviously inferior.

# Qualitative Research Methods

- **Query and Critique**

- Ask people to provide judgements or feedback on a design
- Excellent for getting initial and continued feedback on a design
- May not identify all problems that will actually occur
- Some methods are time consuming

- **Observe Users**

- Observe users actually using the system or prototype
- Typically quick and easy to do
- Some methods can reveal what a person is thinking

# The Introspection Method

- Designer/developer tries the system (or prototype) out
  - Does the system “feel right”?
- Pros:
  - Can notice some major problems in early versions during every day use
- Cons:
  - Not reliable and completely subjective
  - Not valid as introspector is a non-typical user
  - Intuitions and introspection are often wrong
  - Observations may be biased

# Review Based

- Literature review to support/refute design elements
- Pros:
  - Uses huge prior base of knowledge
  - Saves having to redo established experiments
  - Allows cross-study comparisons
- Cons:
  - Results of a study may not apply to other situations

# Design Critique

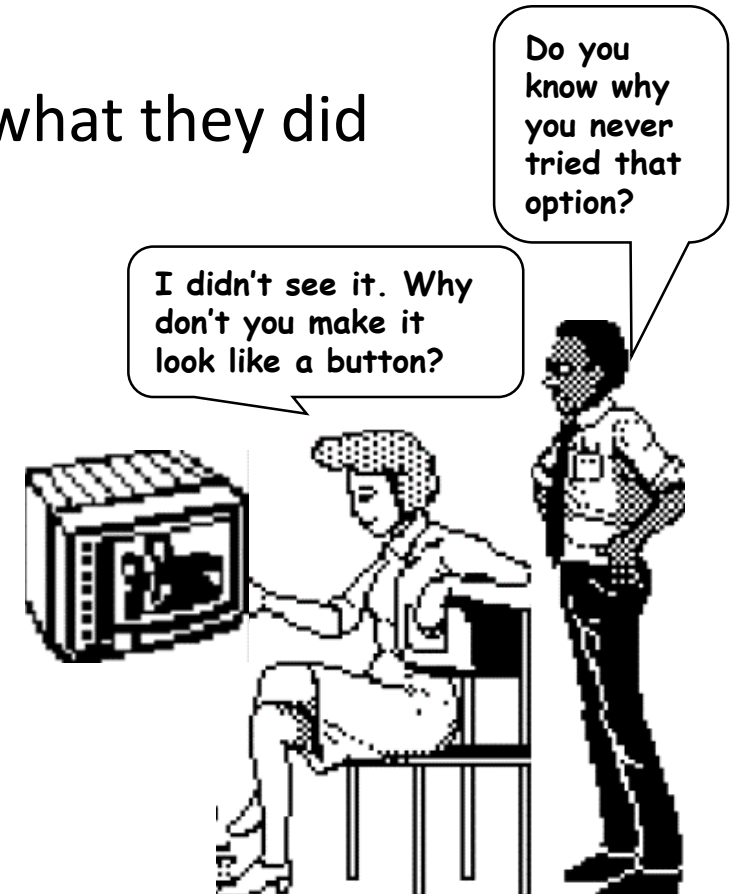
- Useful constructive feedback from a design
  - Collaborative deconstructions of what makes a design successful
  - And what makes it fail...
  - Often used in a studio setting
- Pros:
  - Two-way: designer can provide rationale
  - Critic and critique can engage deeply
  - Can articulate new design possibilities
- Cons:
  - Can become abusive
  - May not involve real users

# Interviews

- Interview users / participants after they have used the system
- Pros:
  - Excellent for pursuing specific issues
  - Often leads to specific constructive suggestions
- Cons:
  - Accounts are subjective
  - Time consuming
  - Evaluator can easily bias the interview
  - User's reconstruction may be wrong

# Retrospective Testing

- Interview to clarify events that occurred during system use
  - Perform an observational test
  - Create a video record of it
  - Have users view the video and comment on what they did
- Pros:
  - Excellent for grounding a post-test interview
  - Avoids erroneous reconstruction
  - Users often offer concrete suggestions
- Cons:
  - Time-consuming
  - Users may forget context





# Questionnaires & Surveys

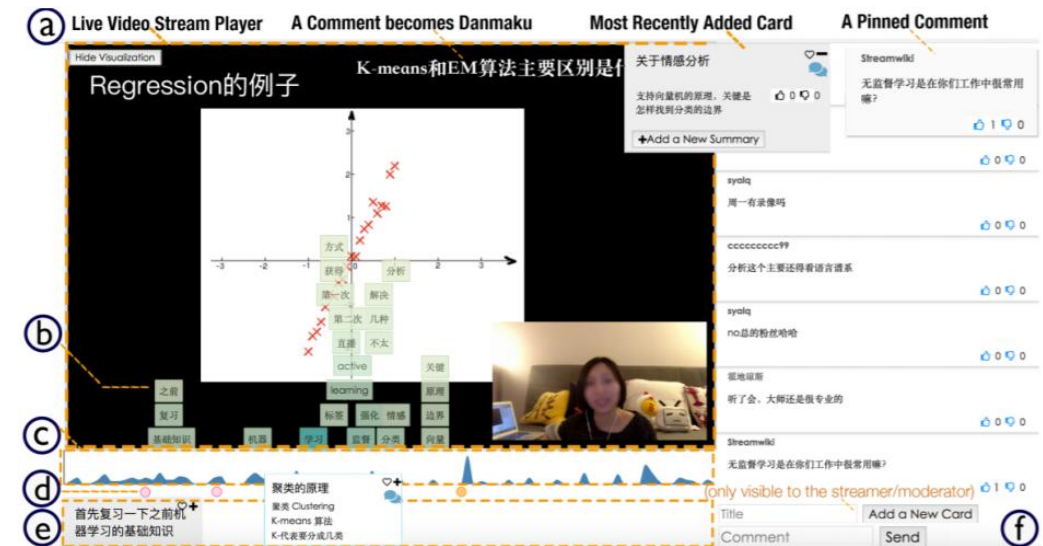
- Set of questions user answer on their own
- Pros:
  - Maintenance / analysis is easy
  - Can reach a wide subject group (e.g. mail / web)
  - Does not require presence of evaluator
  - Results can be quantified
- Cons:
  - Preparation can be time-consuming
  - Only as good as the questions asked
  - Users may misunderstand, or not put in effort

# Qualitative Research Methods

- Query and Critique
  - Ask people to provide judgements or feedback on a design
  - Excellent for getting initial and continued feedback on a design
  - May not identify all problems that will actually occur
  - Some methods are time consuming
- **Observe Users**
  - Observe users actually using the system or prototype
  - Typically quick and easy to do
  - Some methods can reveal what a person is thinking

# Conceptual Model Extraction

- Show the user static images of:
  - Paper prototype, screen snapshots, actual system screens during use
- Have the user try to explain
  - What all elements are
  - What they would do to perform a particular task



# Conceptual Model Extraction

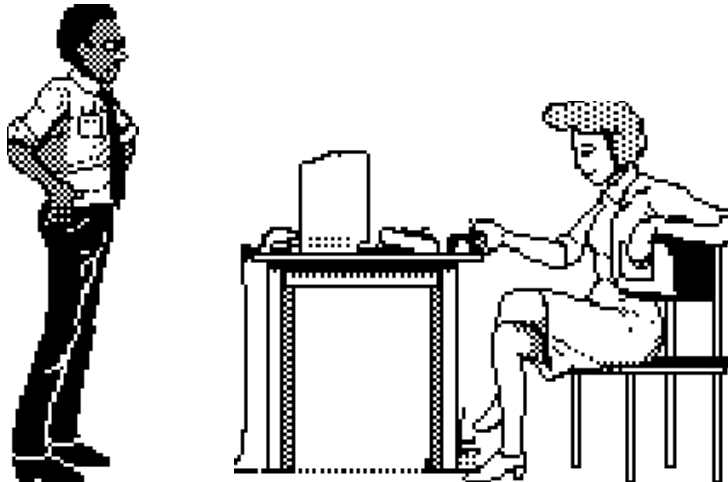
- Initial vs formative conceptual models
  - **Initial:** how person perceives a screen the very first time it is viewed
  - **Formative:** how person perceives a screen after the system has been used for a while
- This approach is:
  - Excellent for extracting a novice's understanding of system
  - Poor for examining system exploration and learning
  - Requires active intervention by evaluator, which can get in the way

# Direct Observation Methods

- Evaluator observes and records users interacting with design/system
  - In the Lab: User asked to complete a set of pre-determined tasks
  - In the Field: User goes through normal duties
- Excellent at identifying gross design/interface problems
- Validity/reliability depends on how controlled the environment is
- Three typical approaches:
  - Silent Observer, Think-Aloud, Constructive Interaction

# Direct Observation: Silent Observer

- User is given the task, and evaluator just watches the user
- Problem
  - Does not give insight into the user's decision process or attitude
  - User's can become frustrated if stuck on a task



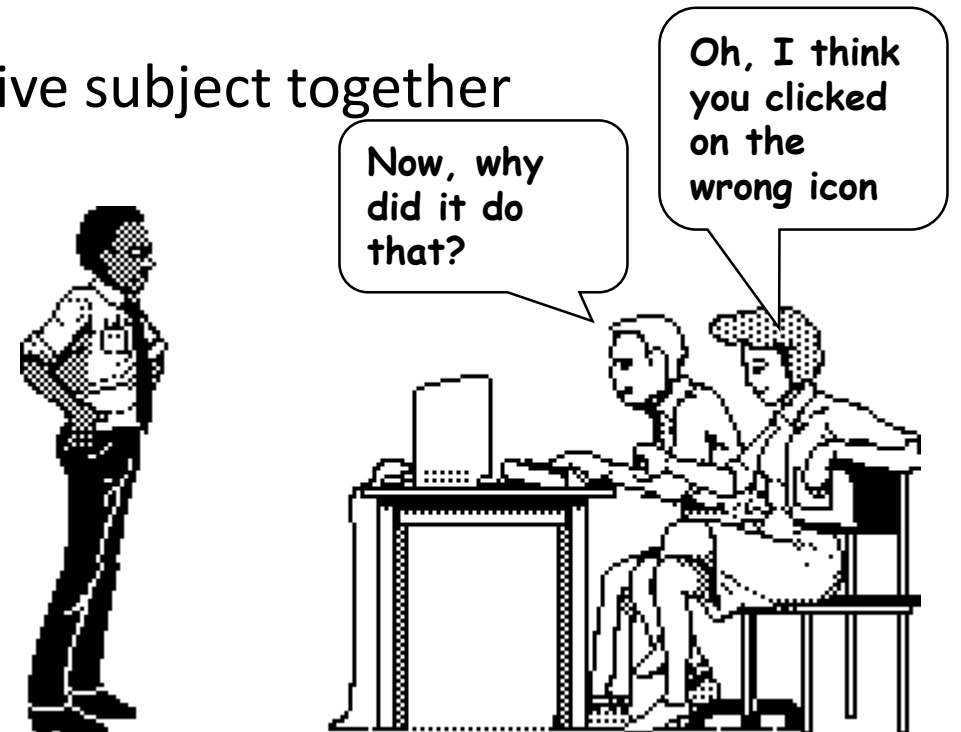
# Direct Observation: The Think Aloud Method

- Subjects are asked to say what they are thinking/doing
  - What they believe is happening
  - What they are trying to do
  - Why they took an action
- Gives insight into what the user is thinking
- Problems
  - Awkward/uncomfortable for subject (thinking aloud is not normal!)
  - “Thinking” about it may alter the way people perform their task
  - Hard to talk when they are concentrating on problem
- Most widely used evaluation method in industry



# Direct Observation: Constructive Interaction

- Two people work together on a task
  - Normal conversation between the two users is monitored
    - Removes awkwardness of think-aloud
  - Variant: Co-discovery learning
    - Use semi-knowledgeable “coach” and naive subject together
    - Make naive subject use the interface
- Results in:
  - Naive subject asking questions
  - Semi-knowledgeable coach responding
  - Provides insights into thinking process of both beginner and intermediate users





# Direct Observation: Recording Observations

- Paper and pencil
  - Evaluators record events, interpretations, and extraneous observations
  - Hard to get detail (writing is slow)
- Audio recording
  - Good for recording talk produced by thinking aloud/constructive interaction
  - Hard to tie into user actions (i.e. what they are doing on the screen)
- Video recording
  - Can see and hear what a user is doing
  - One camera for screen, another for subject (picture in picture)
  - Can be intrusive during initial period of use
  - Companies often build “usability labs” with one-way mirrors, video cams, etc.
- Software Logs
  - Log and time-stamp key events that can be later analyzed
  - May require modification of the software being tested

# Continuous Evaluation

- Case Studies, Field Studies
  - Careful study of “system usage” at the site
  - External observer monitors behavior
- Pros:
  - Good for seeing “real life” use
  - Actual tasks, actual users, actual environments
- Cons:
  - Time consuming
  - Low sample size
  - Can only test situations that arise



Stockholm  
University  
Mobile Life Vinn Excellence Centre  
Stockholm University

## ***Apple Watch in vivo***

Professor Barry Brown  
Mobile Life Research Centre  
University of Stockholm, Sweden

Stefania Pizza, Barry Brown, Donald McMillan, and Airi Lampinen. Smartwatch in *vivo*. CHI '16



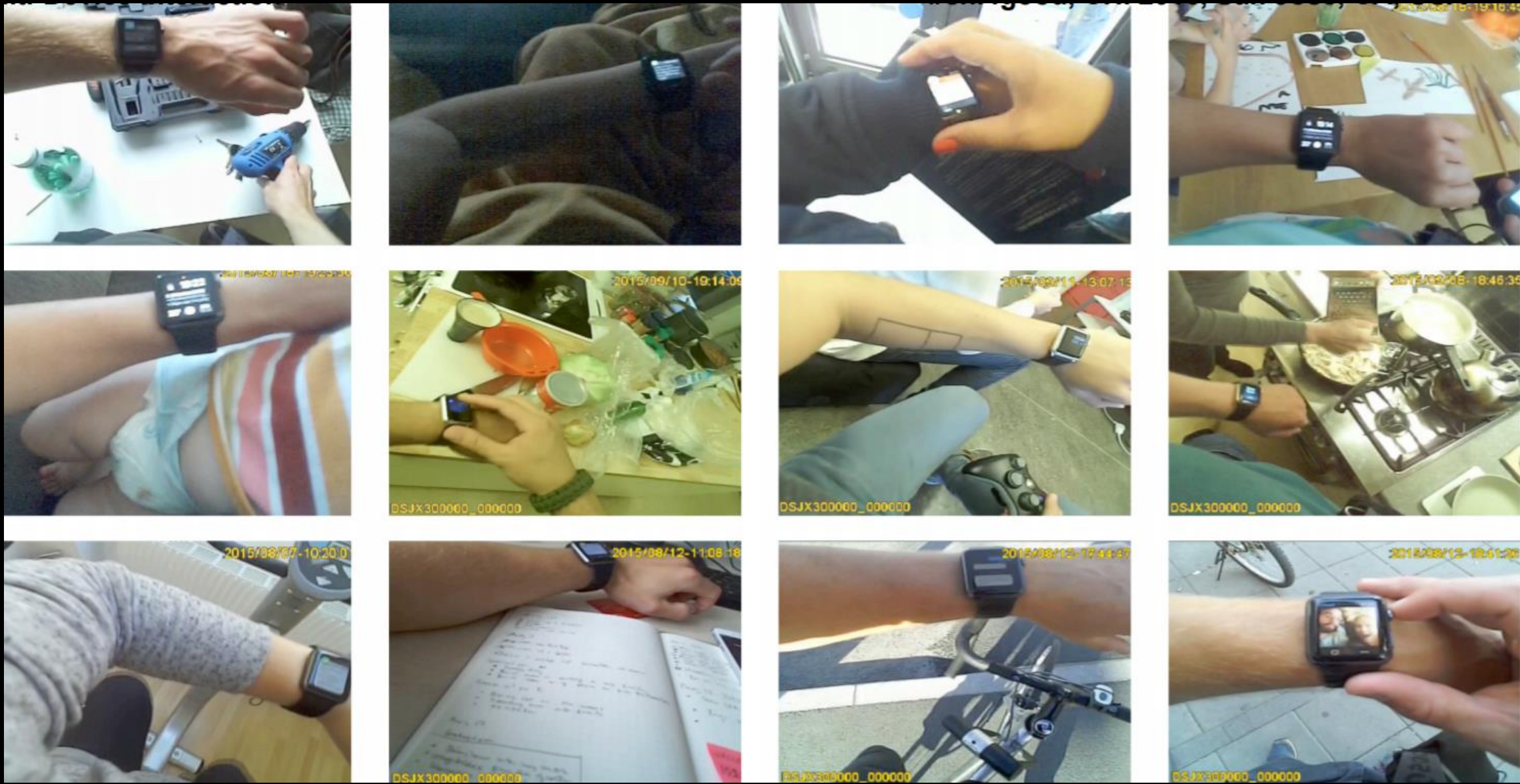
Stefania Pizza, Barry Brown, Donald McMillan, and Airi Lampinen. Smartwatch in *vivo*. CHI '16





Stefania Pizza, Barry Brown, Donald McMillan, and Airi Lampinen. Smartwatch in *vivo*. CHI '16





Stefania Pizza, Barry Brown, Donald McMillan, and Airi Lampinen. Smartwatch in *vivo*. CHI '16

# Getting Situated and Contextualized Data

- To understand participants' experiences natural settings, in their words without having an observer
- It can be collected through
  - Diary Study
  - Experience Sampling

# Diary Study

- Participant-driven data collection: participants write a diary about their experience.
- Self-report at regular intervals, or when an event of interest occurs
- Participants create logs of activities answering given questions.



# Diary Study

- Pros:
  - Non-intrusive, non-interrupting
  - Potential for detailed answers
- Cons:
  - Rely on participants' motivation, effort, and memory
  - May lose many events if participants forget to log

# Experience Sampling

- Experimenter-driven data collection: participants are notified when they need to report their experiences.
  - E.g., using a beeper, phone messages, or custom apps.
- Though sampling
  - Random paging to determine “mental content”
  - Experience sampling: Likert scales or multiple choice questions to assess the context of experience
  - Descriptive Experience sampling: qualitative and open-ended questions.

# Experience Sampling

- Pros
  - Less likely to misremember behavior
  - Collect data anytime with no observer needed
- Cons
  - Interrupting and inconvenient
  - May miss important experiences
  - Potential for post-hoc rationalization

# DRAGON

A Direct Manipulation Interface  
for Frame-Accurate In-Scene  
Video Navigation

Thorsten Karrer, Malte Weiss, Eric Lee, and Jan Borchers.  
DRAGON: a direct manipulation interface for frame-accurate in-scene video navigation. CHI '08.

# Design Project Proposal

- On Oct 1
- Each team will present for 15 minutes + 5-minute discussion
- You're presenting
  - Refined problem + research question
  - Related work
  - Suggested solution
  - Plans for evaluation

# Acknowledgements

- Some of the materials are based on materials by
  - Tovi Grossman, Univ. of Toronto
  - Juho Kim, KAIST
  - Jonathan Lazar, Jinjuan Heidi Feng, Harry Hochheiser,  
Research Methods in Human-Computer Interaction, 2<sup>nd</sup> Ed.

Thank you!