Guide to Designing a Successful Comparative Structured Observation Study

Wendy E. Mackay
mackay@lisn.fr
Université Paris-Saclay, CNRS, Inria
Laboratoire Interdisciplinaire des Sciences du Numerique
Orsay, France

Joanna McGrenere
joanna@cs.ubc.ca
University of British Columbia
Computer Science
Vancouver, Canada

ABSTRACT

Comparative Structured Observation is a qualitative method for HCI (Human-Computer Interaction) researchers and User Experience (UX) designers to assess and advance mid- or late-phase interaction designs. The goal is to obtain rich, qualitative insights from users as they compare design variants to evolve a proposed design concept. Participants are encouraged to consider specific trade-offs between different design variants and help designers assess whether the design is on the right track.

The method takes advantage of structure of a controlled experiment to ensure that participants experience similar, ecologically valid tasks with each variant which enhances their ability to reflect the advantages and disadvantages of each.

This guide is designed to help both HCI researchers and UX designers learn and apply this method. We identify the criteria for creating a successful *Comparative Structured Observation* study, followed by a checklist and an example of a *Comparative Structured Observation* study design. We then discuss how HCI paper reviewers can determine whether or not a particular study should be considered a *Comparative Structured Observation* and discuss how UX designers in industry may benefit from the method. Note that this guide is intended as a practical accompaniment to (Mackay & McGrenere, 2025), published at ACM/Transactions on Computer-Human Interaction.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods.**.

KEYWORDS

Comparative Structured Observation, Design methodology, Interventionist, Mixed-methods, Qualitative methods, Quantitative methods, Research methodology

1 INTRODUCTION

A key creative aspect of HCI research and UX design is to generate novel design concepts. Although we have many methods for finding out about users and generating ideas during the early phases of design, and both qualitative and quantitative methods for evaluating late-phase designs, we lack methods for exploring the middle phase of a design, when the goal is not to determine if the details of the design are correct, but rather to see if the design itself is on the right track. Greenberg and Buxton [5] highlight the difference between *getting the design right* — perfecting a particular design — and *getting the right design* — determining which design direction makes the most sense. The latter goal is not to validate a final design, but rather to obtain insights from users to help verify the design direction and refine the concept.

Mackay & McGrenere [6] address this challenge by introducing *Comparative Structured Observation*, a rigorous mixed-method approach for assessing mid-phase designs by asking study participants to compare and contrast their experiences with different design variants. This guide offers a step-by-step approach to designing and running an effective *Comparative Structured Observation*. The primary audience includes HCI researchers who would like to assess a novel design concept — not from the perspective of performance, but rather to see if a novel concept for an interactive system makes sense to pursue. The method is also appropriate for UX designers who want to quickly understand users' reactions to a novel design direction. Finally, we hope to help HCI reviewers to assess whether or not requirements of a *Comparative Structured Observation*.

2 CHARACTERIZING COMPARATIVE STRUCTURED OBSERVATION

The key goal of *Comparative Structured Observation* is observe participants as they experience selected design variants and to capture their reflections in order to generate implications for design. This includes identifying which aspects of the design do or do not work well and why, as well as providing suggestions for how to improve the design.
We define ***Comparative Structured Observation*** as:

> an interventionist, qualitative method for assessing and advancing a design concept where researchers observe participants as they compare and reflect deeply upon their experiences with selected design variants, exposure to which is structured according to experimental design principles.

We chose the name *Comparative Structured Observation* because it includes the three key characteristics of the method: comparison, structure and observation. We often abbreviate it to *CSO* and can refer to it as a *Comparative Structured Observation* study or just "CSO".

CHI'25, April 2025, Yokohama, Japan
2025. This is the author's version of a handout presented as part of a course at the ACM/CHI'25 conference. Some material is taken from the definitive Version of Record, which was published in *TOCHI*, https://doi.org/10.1145/nnnnnnn.nnnnnnn.

3 *COMPARATIVE STRUCTURED OBSERVATION* CHECKLIST

How should you consider whether or not your design is suitable for being assessed with a *Comparative Structured Observation* study? The following checklists should help you to decide, first, whether or not a *Comparative Structured Observation* study is appropriate for your research question and then what to include in your study protocol and suggestions for how to capture and analyze the data, what to include when you report your findings.

Every *Comparative Structured Observation* study asks study participants to interact with at least two design variants with equivalent activities/tasks to provide them with recent, comparable experiences that enhance their ability to reflect on and compare those design variants. Each design variant must be sufficiently well developed that users can obtain a realistic idea of how they would use each in a real-world context. In experimental design terms, this means that the study must always include at least one *within-participant factor*.

The simplest form of a *Comparative Structured Observation* asks participants to experience two design variants, whereas more elaborate study designs can include additional design variants, additional activities, or both. Activities may include prescribed tasks, steps in a scenario, or other activities relevant to the proposed use of the design. In experiment terms, each study includes at least one *design variant* factor with two or more distinct design variants. Studies may also include additional factors, such as different types of tasks, levels of difficulty or types of expertise. Each of the above requires an associated activity, where the combination of one design variant and one activity is called a *condition*. The activities associated with each design variant should be as equivalent as possible, ideally isomorphic, so that each design variant can be associated with a similar activity.

The following checklists will help you decide if your project appropriate for being assessed with a *Comparative Structured Observation* study.

3.1 Is a CSO study appropriate?

- □ Clearly defined target user population
- □ Design concept based on formative research with users:
  - □ Preliminary user studies
  - □ Research literature
  - □ Both
- □ Design variant is an interactive prototype
  - □ Advanced *Wizard of Oz*
  - □ Medium-fidelity prototype
  - □ Hi-fidelity prototype
- □ One or more *comparison* design variant(s), including
  - □ *Status Quo* (existing system)
  - □ Novel alternative design
  - □ Mix of the above
- □ Grounded activity for participants to perform, such as
  - □ Specific task to accomplish
  - □ Scenario to follow
  - □ Experience within a defined interactive
- □ Equivalent activities for each condition, such as
  - □ Duration
  - □ Number of actions
  - □ Level of difficulty
- □ Ecological validity (Relevance to real-world user context)
- □ Access to members of the target population in a specific setting, such as
  - □ Lab setting
  - □ Field setting
  - □ Combination

3.2 What belongs in the study protocol?

If your project has met the above requirements, you can design your study protocol. Ensure that your research or design questions are not framed as causal hypotheses, but rather relate to user's reflections based on their experiences with each design variant.

- □ Choose the design variants
  - □ Design variant 1 + Design variant 2
  - □ Design variant 1 + Status quo variant
  - □ Design variant 1 + [status quo] and/or [mulitple variants]
- □ Specify additional factors (optional)
  - □ Multiple task types, e.g. search vs. decide
  - □ Level of expertise, e.g. novice vs. expert
  - □ Level of difficulty, e.g., easy, medium, hard
- □ Calculate number of conditions
  - □ # design variants x # other factors (optional)
- □ Create equivalent activities for each condition
  - □ Task
  - □ Scenario
  - □ Experience
  - □ Sequence of activities
  - □ Other
- □ Calculate # conditions per participant
- □ Allocate design variants and activities to participants, ideally counter-balanced for order

Data collection

- □ Audio recording
- □ Video recording
- □ Hand-written notes
- □ Participant action log
- □ Rank of design variants
- □ Questions about variants
- □ Quantitative data, e.g. performance [optional]

Design the study protocol

- □ Brief participants about the study
- □ Obtain informed consent
- □ Demonstrate design variants (optional)
- □ Practice using design variants (optional)
- □ Part 1: For each condition (design variant + activity)
  - □ Perform activity
  - □ Answer questions about design variant
- □ Additional parts (optional)
- □ Open-ended exploration (optional)
- □ Reflective interview to compare design variants: pros, cons, trade-offs

3.3 Data Analysis

- □ Thematic analysis
  - □ Transcript of talk aloud during activities
  - □ Transcript from interview
- □ Log analysis
- □ Question analysis

3.4 Findings

- □ Specific pros/cons/trade-offs for each variant
- □ Key inductive themes
- □ Potential improvements
- □ Potential new direction
- □ Overall assessment of design variants

4 EXAMPLE

HCI researchers have already published a number of variations of *Comparative Structured Observation* in the HCI literature, although not all of them have been labeled as such. The following example illustrates how to apply the *Comparative Structured Observation* method to evaluating a novel design concept.

4.1 Example: HelpCall

Concept: Oder adults commonly depend on their family or social circle for remote tech support. However, standard video-conferencing systems are inadequate for their needs. *HelpCall* is a video-conferencing add-on that captures, displays and saves steps from live demonstrations given by remote family members to help older adults learn computer tasks. The designers want to assess two design variants — Tooltip and List — to see if the *HelpCall* concept can successfully mediate these remote video calls.

Is a CSO study appropriate? Yes

- *User population:* Older adults
- *Formative research:* Literature review + cognitive walkthrough with older adults and design experts
- *Interactive prototype:* Interactive mock-up of predetermined tasks + Wizard-of-Oz control
- *Comparison variants:* Two: Tooltip + List + Status Quo
- *Activities:* six learning tasks, beginner and intermediate difficulty
- *Equivalent activities:* three equally unfamiliar tasks per participant
- *Ecological validity:* Mimics real video call between an older adult and a helper
- *Access:* Older adults in their own homes with a simulated video call

What belongs in the study protocol?

- *Design variants:* Tooltip + List + Plain video call (status quo)
- *Additional factors:* None
- *Conditions:* Three
- *Equivalent activities:* Six possible tasks, chosen for lack of familiarity
- *Conditions per participant:* Three
- *Counter-balance order:* Status quo first, then counter-balanced order for two design variants
- *Data collection:* Video recordings, screen recordings, comparison questionnaire, notes
- *Protocol:* Briefing, Informed Consent, three tasks, three questionnaires, comparative interview, debriefing
- *Data analysis:* Mixed thematic analysis

Data analysis

- *Qualitative:* Mixed thematic analysis of comparative interview transcripts combined with behavior observed during tasks
- *Quantitative:* Six metrics: pre-task self-rated confidence in completing the task (1-5 Likert scale), post-task confidence with and without augmented assistance, duration of the helper demonstration and no-helper attempt rounds and error rate.

5 *COMPARATIVE STRUCTURED OBSERVATION* FOR UX DESIGNERS

Although Mackay & McGrenere explicitly targeted HCI researchers, a simplified version is also appropriate for product designers who need an efficient method for assessing a mid-phase design concept for a new or revised product. UX designers are often under tighter timelines than academics but face less stringent requirements for rigor in their user studies. They may thus find *Structured Observation* useful, but explicitly relax certain requirements, such as not fully counter-balancing task order; or restricting their thematic analysis to key themes, such as "breakdowns, workarounds and user innovations" [7]. However, even a "relaxed *Comparative Structured Observation*" will let participants experience and reflect on design variants, thus providing designers insights about complex new design concepts.

ACKNOWLEDGMENTS

This work was partially supported by European Research Council (ERC) grant № 321135 "CREATIV: Creating Co-Adaptive Human-Computer Partnerships" and Natural Sciences and Engineering Research Council of Canada (NSERC) grant № RGPIN-2017-04549 "Highly personalized user interfaces".

RECOMMENDED READING

[1] Hugh Beyer and Karen Holtzblatt. 1997. *Contextual Design: Defining Customer-Centered Systems.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[2] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

[3] Thomas D Cook and D T Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Houghton Mifflin, New York.

[4] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[5] Saul Greenberg and Bill Buxton. 2008. Usability Evaluation Considered Harmful (Some of the Time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08).* Association for Computing Machinery, New York, NY, USA, 111–120. https://doi.org/10.1145/1357054.1357074

[6] Wendy Mackay and Joanna McGrenere. 2025. Comparative Structured Observation. *ACM Trans. Comput.-Hum. Interact.* 32, 2 (jan 2025), 27. https://doi.org/10.1145/3711838

[7] Wendy E. Mackay. 2023. *DOIT: The Design of Interactive Things. CHI 2023 Preview.* Inria, Paris, France.

[8] Anselm L. Strauss and Juliet M. Corbin. 1998. *Basics of qualitative research: techniques and procedures for developing grounded theory.* Sage Publications, Thousand Oaks, California. 312 pages.