Comparative Structured Observation

WENDY E. MACKAY, Université Paris-Saclay, CNRS, Inria, France
JOANNA MCGRENERE, University of British Columbia, Canada

Although HCI researchers often generate and compare new design concepts, they lack an established method for rigorously conducting *qualitative assessments*. We define and characterize *Comparative Structured Observation* as a qualitative research method that takes advantage of the structure of controlled experiments to generate comparable, ecologically relevant experiences with two or more design variants, often implemented as medium-fidelity prototypes. Researchers observe users and ask them to compare and reflect on each variant. We identify criteria for creating a successful *Comparative Structured Observation* study and illustrate variations of the method by analyzing four published studies. We also examine six additional studies (three "near" and three "far") to clarify the boundary between what should and should not be considered a *Comparative Structured Observation*. We discuss the benefits and limitations of the method and argue that gathering comparative reflections about design variants can help researchers assess and advance their design concepts.

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

Additional Key Words and Phrases: *Comparative Structured Observation*, Design methodology, Interventionist, Mixed-methods, Qualitative methods, Quantitative methods, Research methodology

1 INTRODUCTION

A key role of Human-Computer Interaction (HCI) research is to generate novel concepts that inform the design of interactive systems. HCI researchers can *generate* new design insights by drawing upon a variety of well-defined empirical qualitative methods in the early phases of a design-based research project. For example, naturalistic observation helps researchers understand user needs in field settings and produce implications for design; whereas participatory design involves target users and researchers working together to explore new design concepts.

However, once a design concept begins to take shape, often in the form of a medium-fidelity prototype, it remains challenging to assess it with users. The goal at that stage is not yet to validate a final design, but rather to obtain insights from users that can inform the design direction and refine the concept. We are interested in clearly defined qualitative methods that take advantage of users' reflections to assess and advance design concepts as HCI researchers explore, develop and assess them. Of course HCI researchers derive design ideas from a variety of sources, including theory [3, 45] and conceptual work [26, 49]. Even so, regardless of the idea's origin, researchers must still find a way to empirically assess whether or not it offers a promising design direction.

Researchers from Psychology and Computer Science came together in the early 1980s to found the CHI (Computer-Human Interaction) conference and its community. Social scientists were encouraged to demonstrate not only their insights about users but also the corresponding implications for design. Similarly, technically trained researchers were encouraged to not only design novel systems but also demonstrate positive benefits for users. This combination of

Manuscript submitted to ACM

Psychology and Computer Science approaches remains central to the CHI community even as the field has expanded to encompass additional social sciences, engineering disciplines, and design fields.

This historical grounding has led some HCI researchers (and reviewers) to insist that new system designs must be compared to other designs with controlled experiments and quantitative measures. However, such experiments are only appropriate for evaluating clearly articulated, testable hypotheses, such as demonstrating improved performance with a specified interaction technique. Such studies are far less relevant when the researcher's goal is to assess the user's subjective reactions to potential design variants or to choose among alternative design directions.

Some HCI reviewers assume that evaluating a novel system requires quantitative methods, which has in turn led some authors to avoid qualitative study designs, even when quantitative assessments are either too early or altogether inappropriate. Greenberg and Buxton [25] argue that this emphasis has a cost, since it encourages HCI researchers to focus on *getting the design right*, i.e. selecting a design early and perfecting it, rather than *getting the right design*, i.e. generating and comparing multiple design alternatives to explore their strengths and weaknesses. Prominent HCI researchers have gone further, speculating that HCI's over-emphasis on quantitative usability evaluations may be linked to the low adoption of published interaction design research outside of the field [25, 32, 43].

Of course current HCI research includes a variety of accepted qualitative methods. For example, design-oriented researchers who try to "get the right design" use qualitative methods to critique ideas early in the design process. Industry-based user experience (UX) designers who want to "get the design right" often run usability studies to gather customer reactions about near-final designs. However, researchers who have already invested significant time and effort in developing interactive prototypes face a dilemma if they want to *get the right design*:

Is it better to run an informal qualitative study that produces "fuzzy" claims about how much users "like" the new design or to run an inadequately controlled experiment with hypotheses chosen more to demonstrate statistical significance than for their relevance to assessing the design concept? Should the authors add an experiment to appease reviewers even if the primary contribution stems from qualitative insights? What about reviewers who are excited by a new concept but reject the article because the inadequacy of the experiment detracts from the work?

Like many of our colleagues, we have been frustrated by the lack of an accepted, named method with well-articulated criteria that bridges the gap between informally testing an early design prototype and formally testing claims about causal relationships among design variants. We have each struggled with these issues as our respective research foci have evolved, shifting from primarily quantitative experiments to qualitative and mixed-methods studies. Over the years, we have expanded our use of quantitative methods to include an increasing emphasis on qualitative elements. We have published multiple studies that combine the underlying rigor of controlled experiments that manage the systematic presentation of design variants with the richness of qualitative data analysis. This has often led to a disconnect, where some reviewers compliment the rigor of our methodology while others focus on the "missing" quantitative data. We have heard the same challenge from colleagues who conduct research similar to ours. We thus argue that HCI needs to explicitly characterize a well-defined qualitative method for assessing promising design concepts as they emerge from exploratory research, without having to fully implement and validate the final design. Our goal is to provide a commonly agreed-upon set of guidelines for rigorously conducting such studies.

Just as Braun & Clarke [7, 8] demarcated *Reflexive Thematic Analysis* in psychology, we define and characterize *Comparative Structured Observation* as a method for structuring qualitative studies where users explicitly compare and reflect upon their experiences with specified design variants.

To arrive at this characterization, we reflected deeply on our own mixed-methods published research and that of others who have published similar studies. We abstracted the key methodological characteristics that produce a rigorous qualitative comparison. This characterization required multiple iterations, both to establish which characteristics are essential and which are optional, as well as to clearly articulate each characteristic. We sought informal feedback on drafts from approximately a dozen researchers with backgrounds in the Psychology+CS approach to HCI.

Comparative Structured Observation takes advantage of experimental (or quasi-experimental [14]) protocols to facilitate comparable conditions with respect to design variants. We argue that the key is *comparison*: *Comparative Structured Observation* provides users with comparable experiences, either with multiple new design variants or with the status quo, which lets them reflect deeply on the advantages and disadvantages of each. Researchers observe and compare how users experience and react to each variant and probe for additional insights. Organizing these experiences according to experimental design principles also helps to account for nuisance factors such as learning and fatigue, since users can reflect on how their experiences change under these different conditions.

After first characterizing *Comparative Structured Observation*, this article analyzes four of the published studies that we used to extract and refine the characteristics. We further analyze additional published studies to clarify the boundary between what might or might not be considered a *Comparative Structured Observation* and establish clear criteria for conducting a rigorous study. We hope that providing this common foundation, rationale and suggested best practices will facilitate adoption of this method by HCI researchers, not only by helping them to design and conduct more effective comparative qualitative studies but also by assessing such studies fairly and consistently. Note that our primary audience for using this method are HCI researchers with prior training in both qualitative and quantitative methods. However, we also believe that practitioners trained in *Comparative Structured Observation* methodology will also be able to use it productively in their design practice.

2 CHARACTERIZING COMPARATIVE STRUCTURED OBSERVATION

The primary goal of *Comparative Structured Observation* is to **capture insights from participants' reflections about selected design variants in order to generate implications for design**, either by identifying the aspects of the design that do or do not work well and why, or by providing direction for how to improve the design. **Comparison is essential** — both participants and researchers must be able to compare and reflect on the advantages and disadvantages of each variant relative to its context of use. Because *Comparative Structured Observation* often yields knowledge about how tasks moderate the success of different design choices, sometimes characterized as *What design works when? how? and why?*, design implications should be richly contextualized.

The next section defines *Comparative Structured Observation*, with a rationale for the name. We then identify the specific characteristics that constitute a *Comparative Structured Observation* study. (Section 6.1 describes the history of how we came to adopt *Comparative Structured Observation* as a design method, which provides context for the definition and characteristics.)

2.1 Definition and Rationale for the Name

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 evokes key desired characteristics of the method. It emphasizes qualitative observation of how study participants experience and compare design variants, with structured tasks or events that are usually imposed by the researcher. We note that the shorter term *Structured Observation* does appear in a number of qualitative research fields. For example, sociology textbooks [19] refer to *Structured Observation* as a naturalistic observation conducted in a systematic way, with no explicit intervention by the researcher. *Structured Observation* also appears in several applied social sciences, including educational research [16], organizational behavior [38, 41], and healthcare [11], who often use "systematic observation" [17] as a synonym. However, these publications use the term *Structured Observation* informally, mostly to remind researchers that creating a structured observation checklist will facilitate subsequent data analysis. We feel that the longer term *Comparative Structured Observation* both captures what we mean and will not be confused with prior and informal uses of the shorter term *Structured Observation*.

Our definition of *Comparative Structured Observation* is explicitly interventionist, since the researcher imposes design variants for the study participants to experience, either by generating tasks for them to perform or ensuring the presence of naturally occurring events.

Comparative highlights the centrality of comparison, which is key to the design of controlled, quantitative experiments [14]. Here, we obtain similar benefits with qualitative studies, since exposing participants to two or more design variants encourages participants to reflect on and compare the strengths and limitations of each. Participants explicitly compare their experiences with ecologically grounded tasks which provides a rich foundation for critiquing the proposed design elements, with a correspondingly richer, deeper feedback. Comparison also helps researchers assess the impact of design decisions more clearly, especially those that vary across the design variants.

Structured refers to the structuring of study participants' activities and the ordering of their exposure to different design variants, e.g., through counter-balancing. The researcher selects specific tasks or events to facilitate both the participant's and the researcher's ability to compare and reflect on the variants. The tasks should be grounded in real-world context to be as ecologically valid as possible. The researcher takes advantage of the principles of experimental design to control the presentation of conditions (the design variants) and tasks as well as to explicitly account for various "threats to validity" [14, 48], such as primacy and recency effects, learning and fatigue effects, and confirmation bias [28]. Note that borrowing the structure of controlled or quasi-experiments [14, 48] does not imply that we also borrow the corresponding emphasis on quantitative measures, nor the expectation that the studies should be designed to test operationalized hypotheses or establish causal relationships.

Observation highlights researcher's observation and recording of qualitative aspects of the users' behavior. Researchers and participants must interact with each other, usually in the form of a post-hoc interview, in order to gather detailed reflections about participants' in-context experiences with the system of interest. Note that interviews alone are insufficient to be considered "observation"—in all cases, researchers need an independent source of data that captures each participant's behavior with respect to the design. This combination of observation and interviews aligns well with the mixed-method framework advocated by Johnson & Onwuegbuzie [28], where interviews conducted during or after observing how participants have performed tasks or experienced events add to the richness of the qualitative results that can be drawn from the study.

2.2 Characteristics of *Comparative Structured Observation*

We characterize *Comparative Structured Observation* as follows:

Researchers explicitly define at least two comparable design variants that all participants experience as they perform the same or equivalent tasks. In experimental design terms, this means that the study must always include at least one within-participant factor. Design variants may include:

(1) two or more variants of a novel design concept, e.g., designs for synchronous online lectures that add different types of simulated attendee proximity cues.
(2) a new design and a baseline, where the baseline can be a different system, e.g., a novel vs. an established document editor, or the absence of the new system, e.g., a browser as the baseline vs. the browser with a novel plugin.
(3) two or more variants of a novel design concept and a baseline, which simply combines (1) and (2) into a single study.

Design variants are usually expressed as medium-fidelity prototypes, although low-fi and hi-fi prototypes are possible. Designs must be sufficiently mature to capture how users will experience different aspects of the design concept through performing relevant tasks.

Researchers derive participant tasks based on formative research, either conducted directly by the researcher or derived from the research literature. Tasks are chosen to maximize ecological validity and generate qualitative data relevant to the design concept. When imposing tasks is difficult or irrelevant, the basis of comparison may be derived from an activity's inherent structure, e.g., naturally occurring notifications that interrupt the user from a primary task.

Lab and field settings are both appropriate. Field studies offer greater ecological validity, but are not always possible, e.g., when a design implementation is not sufficiently robust for field use. Lab studies may allow participants to experience relevant events more quickly than under field conditions or confront them with rare but important events.

Participant exposure to tasks and design variants meets best practices of experimental or quasi-experimental design to mitigate known threats to validity.

Participants compare and reflect upon their experiences with each design variant. The study protocol includes capturing qualitative data where participants explicitly compare and reflect deeply on their experiences with the design variants.

Researchers compare and reflect on participants' experiences based on direct observation or other sources of rich data. Ideally, observation is live and in person, but may be conducted remotely, e.g., via video link, or involve recorded video or other rich log data.

Researchers conduct post-hoc interviews, either semi-structured or open-ended, after participants have experienced the design variants. These interviews probe for additional information about participants' experiences or what researchers observed during the study, with the emphasis on comparing the different design variants.

Researchers emphasize gathering qualitative data to obtain rich, contextualized insights about design variants. This advances the goal of driving design innovation based on the grounded reflections reported by participants and the researcher's interpretations of their behavior and reflections.

Researchers may also capture quantitative data to help interpret qualitative data, but this is considered secondary. For example, the researcher may collect self-reported measures of satisfaction or measure time to complete a task and compare it to the user's perception of their own performance.

Researchers analyze qualitative data using recognized methods, such as reflexive thematic analysis [8] or grounded theory [50]. Thematic analyses will normally include both inductive and deductive analysis, since the structure of the participants' tasks or experiences will provide themes for some of the "top-down" analysis, and "bottom-up" insights are always welcome.

3 RELATIONSHIP TO OTHER METHODS

For simplicity, we use the abbreviated term *Structured Observation* from this point forward, much as Braun and Clarke's [8] "reflexive thematic analysis" is often simplified to "thematic analysis".

So, how does *Structured Observation* compare to other empirical HCI methods? HCI is a highly multidisciplinary field with methods and analysis techniques drawn from multiple disciplines, not only from the social and natural sciences, but also from design and engineering [3, 36]. Some HCI researchers take an explicitly positivist stance and run controlled laboratory experiments to establish general cause-and-effect relationships. They record quantitative measures, usually performance metrics such as speed and error, and often ask Likert-style questions about participants' preferences. Statistical tests determine if, for example, one interaction technique is better than another according to certain dependent measures, given specified tasks and environments. Such experiments are appropriate when clear, quantifiable hypotheses can be established and ecological validity is deemed less important.

However, many HCI researchers are not interested in quantitative performance measures and instead seek insights about complex, multi-faceted design concepts. Some researchers gather qualitative feedback from experts and other stakeholders, without directly involving users. For example, design critiques [1] are conducted by experienced designers to gather "constructive criticism" about a design to determine whether or not the design meets its objectives. However, other researchers seek rich, contextualized qualitative data directly from the target user population to assess the value of the concept or to inspire new design directions. Such research fits more comfortably within an interpretivist stance, where the goal is not to make strong generalizations, but rather to better understand the user's perspective. HCI researchers can draw from a large, established literature regarding quantitative evaluation methods, but have fewer accepted methods for rigorously structuring user studies that maintain a qualitative focus. This creates a disparity across research goals and assumptions, which can lead to conflicts between some HCI researchers with different disciplinary backgrounds when they attempt to justify their research approaches.

Several HCI researchers have addressed this conflict explicitly. For example, Gray & Salzman's [23] critique of five highly cited HCI articles argues that researchers should follow Cook & Campbell's [14, 48] advice for creating "quasi-experimental designs" to maximize ecological validity while maintaining a level of experimental control given specific "threats to validity". However, as Runkel & McGrath [40, 47] demonstrate, no individual research method is ever fully sufficient, since each method optimizes one result at the expense of others. Mackay & Fayard [37] address this by advocating triangulation across methods, a "mixed methods approach" [28] that helps balance the trade-offs inherent in running quantitative and qualitative studies. They also highlight the differences between the research goals of social scientists who seek better understanding of human behavior, cognition, and emotion; and those of HCI researchers, who are more interested in understanding users in order to generate implications for design.

Table 1 compares and contrasts *Structured Observation* to ten empirical methods with well-defined characteristics commonly used by HCI researchers, including some borrowed from psychology and the social sciences, and some created specifically by HCI researchers. We do not include general methodological approaches, such as participatory design [24], action research [31] and activity theory [10], since they encompass a broad range of specific methods and are less amenable to the precise characterization given in the table. Because many empirical HCI studies can be viewed

Research methods	Source Discipline	Project Phase	Research Setting	Primary Research Goal	Intervention?	Primary Analysis	Comparison Required?	Participant Reflection?
Comparative Structured Observation	Psych, Soc Sci	later	any	design implications	yes	qualitative	yes	yes
Naturalistic Observation	Anthro	early	field	understand users	no	qualitative	-	-
Contextual Inquiry	HCI	early	field	understand users	yes	qualitative	-	yes
Experience Sampling	Soc Sci	all	field	understand users	yes	qualitative	-	yes
Diary Study	Soc Sci	all	field	understand users	yes	qualitative	-	yes
Cultural Probe	HCI, Design	early	field	idea generation	yes	qualitative	-	-
Technology Probe	HCI	later	field	evolve design	yes	qualitative	-	yes
Usability Study	Ergo, HCI	later	any	identify problems	yes	quantitative	-	-
Quasi Experiment	Soc Sci	later	field	causal relationships	yes	quantitative	yes	-
Controlled Experiment	Psych	later	lab	causal relationships	yes	quantitative	yes	-

Table 1. Contrast between *Comparative Structured Observation* and other methods common in HCI research, including source discipline; research phase, setting, and goals; intervention and analysis type, and whether the method explicitly requires comparison and reflection.

as "user studies", we omitted them as too general and poorly defined. Similarly we omitted "field studies" as being an overly general term for user studies conducted in the field.

Naturalistic observation[2] is used extensively in anthropology and other social sciences to collect field data, without intervention or other manipulation by the researcher. The primary goal is to understand naturally occurring human behavior, primarily by analyzing qualitative data, with no explicit comparisons or reflections by participants. HCI researchers often use naturalistic observation to discover user needs and prepare for more highly structured design methods.

Contextual Inquiry [5] is an anthropology-inspired HCI method where researchers interview people in a field setting, usually a corporation, to identify their needs with respect to a new or existing technology. The primary goal is to better understand user needs in context, with intervention in the form of interview questions. Although comparison is not required, study participants are sometimes asked to reflect on their experiences with respect to other systems.

Experience sampling[33] and *Diary studies* [46] are borrowed from the social sciences. They ask participants to record specific experiences over a period of time in order capture qualitative, contextualized user experiences, although quantitative data may also be logged automatically. These studies involve interventions in the form of semi-structured questions and optionally a prototype, but do not involve comparisons, although they do ask participants to reflect upon their experiences.

Speed Dating [18] is a design-oriented interventionist method created to assess design concepts and reflect on their merits by exposing users to multiple lo-fidelity sketches, storyboards or mock-ups of diverse design concepts. This method asks participants to quickly assess a relatively large number of early-stage designs, rather than the two to three that are common in *Structured Observation* with somewhat more advanced designs. Explicitly comparing the design concepts to each other is not required. Exposure to the design variants is not controlled according to the principles of experimental design, nor are tasks required.

Cultural probes [22] are a design-inspired HCI method intended to produce ideas by creating artifacts and assigning evocative tasks that participants perform in specified settings. They do not seek to understand users in the traditional, social science sense, but instead seek to stimulate the designer's imagination. The data is qualitative, with no comparisons, although participants may sometimes be asked to reflect upon their experiences.

Technology probes [27] ask users to try a new technology for a specified period of time in a field setting. This is an explicitly multi-method HCI approach that addresses three different types of research goals: obtaining qualitative and quantitative data about users' experiences; assessing technical constraints relative to the specific use setting; and inspiring ideas that advance the design concept. Technology probes involve intervention, both in the installation of the technology and in the form of post-hoc interviews. Participants are not asked to make comparisons but are asked to reflect upon their lived experiences.

Usability studies [42] are an HCI method inspired by Human Factors or Ergonomics research, with the goal of testing late-stage prototypes or products to determine if and how well users can access features to achieve their goals. The focus can be on either qualitative data about how the interactive technology was experienced or on quantitative performance data. Researchers intervene by asking users to perform pre-defined tasks with the prototypes/products. Although comparison is not required, participants may be asked to reflect on their preferences with respect to the system.

Quasi-experiments [14, 48], or more generally, field experiments, are drawn from psychology and other natural sciences, and seek to determine causal relationships. Although based on controlled experiments, they are designed to address potential threats to validity that arise when experimental conditions cannot be fully controlled in the field. Quasi-experiments focus on quantitative data generated by participants' performance on pre-defined tasks, although qualitative data may also be collected. Comparison is required although participant reflection is not.

Controlled experiments [14, 39] are drawn from psychology and other natural sciences, with the explicit goal of determining causal relationships across factors. Researchers intervene by defining initial hypotheses that can be operationalized into participant exposure to conditions and well-defined tasks, with quantifiable measures of performance. Experiments focus on quantitative data and comparison is required although participant reflection is not.

The above research methods each address different research needs. *Comparative Structured Observation* provides a useful complement to these methods, and borrows characteristics from both qualitative and quantitative approaches. It combines the rigorous design of an interventionist study based on well-established experimental design principles with the careful analysis of participants' comparative reflections using well-known qualitative analysis methods. The goal is to organize the presentation of conditions so as to maximize the participant's ability to systematically compare and reflect upon relevant design characteristics and thus advance the design concept.

4 *COMPARATIVE STRUCTURED OBSERVATION* CRITERIA

Determining whether or not a particular study design should be considered a *Comparative Structured Observation* requires assessing it according to a specific set of criteria. Table 2 presents two checklists: the first is designed to help researchers determine whether or not a particular study design should be considered a *Comparative Structured Observation* study. The second suggests features for designing and conducting a "good" one.

Our goal is to improve the quality of future qualitative studies by encouraging both HCI researchers and reviewers to employ these criteria in a consistent way. Note that each criterion for what constitutes a good use of the method could be expanded into a full set of its own criteria. For example, best practice for reporting qualitative results includes a brief summary of the most "obvious" points, followed by a more detailed analysis of the most important or surprising

Dimension	#	Study qualifies as a Comparative Structured Observation if it ...	Study qualifies as a good Comparative Structured Observation if it ...
Design Concept Basis	1	builds on design concepts influenced by formative research, ideally conducted by the researcher but also from the literature.	reports on substantive formative research, e.g. a well-run participatory design workshop that includes reflection by both participants and researchers.
Role of Comparison	2	ensures that each participant experiences at least two design variants in the study, e.g. different novel designs, variants within a novel design, or a baseline.	chooses design variants that meaningfully advance the design concept(s) and avoids straw-man comparisons.
	3	structures participant activities so they can experience and compare design variants, e.g. perform equivalent tasks with each design variant.	chooses and structures meaningful activities for participants, e.g., ecologically valid tasks.
	4	structures comparisons according to experimental design or quasi-experimental design principles, e.g. counter-balance tasks for order.	justifies the protocol relative to the setting (lab or field) and comparisons being made, according to best experimental or quasi-experimental design practices.
Type of Data Collected	5	records participants' comparisons and reflections on the qualitative differences in their experiences with the design variants, e.g., through interview questions.	includes well-designed interviews or surveys that elicit detailed, thoughtful comparisons by participants after exposure to the design variants.
	6	records participants' interactions with each design variant, e.g. through video recordings or high-quality cinematic logs.	collects rich, in-situ observational data or the best-possible alternative, e.g. remote video or substantive experience samples.
	7	records quantitative data only if it helps add context to qualitative data; e.g. percentage of time participants spent in an activity.	records quantitative data, if relevant, to contextualize qualitative data, e.g. participants' interactions with design elements that clarify their experiences.
Type of Data Analysis	8	analyzes participants' comparisons and reflections about the design variants; e.g. with reflexive thematic analysis.	demonstrates that participants have compared and reflected deeply about their experiences with the design variants.
	9	analyzes researchers' independent assessment of the participants' experiences; e.g. with reflexive thematic analysis.	leverages rich, qualitative data so that researchers can independently assess participants' reflections.
	10	treats qualitative analysis as primary.	conducts and reports a rigorous qualitative analysis according to the best practices of a well-established qualitative method.
	11	treats quantitative analysis as secondary.	analyzes quantitative data according to the best practices of well-established quantitative methods, either or both descriptive or inferential statistics.
Results	12	reports findings and analysis to advance one or more design concept(s).	explicitly discusses the implications for design and how the design concept(s) should evolve, based on the study results.

Table 2. Left: *Comparative Structured Observation* checklist. Right: Suggested features of a "good" *Comparative Structured Observation* study.

findings; and best practice for creating an experimental design includes generating "equivalent" tasks and structuring their presentation to avoid unwanted nuisance effects, such as learning or fatigue. Unpacking each criterion is beyond the scope of the current article, and we expect researchers who use *Comparative Structured Observation* to have sufficient prior training in both qualitative methods, e.g., [7, 8, 12, 15], and quantitative design and analysis methods, e.g., [14, 39].

5 PUBLISHED STUDIES USING COMPARATIVE STRUCTURED OBSERVATION

The following four case studies from the HCI literature meet the criteria of *Comparative Structured Observation*, but could be improved if they followed the suggestions listed in right-hand column of table 2. We chose these articles, including some of our own, to illustrate different variations of the method. Note that Koch et al. [30] and Kahn et

al. [29], our own most recent uses of this methodology, use the term *Structured Observation* in the publication, but not the full term *Comparative Structured Observation*. It is only in writing this article that we feel that the longer term of *Comparative Structured Observation* more accurately captures the method.

Each of the four studies follows a properly controlled experimental protocol and although the authors collect quantitative data, the primary purpose of each study is to collect and analyze qualitative data. Each case study describes the design concept being evaluated, identifies the conditions being compared, and describes the study design, research goals and key findings. They all suggest implications for design, explicitly or implicitly, which designers can use to iterate on the design concept. We conclude with an analysis of how the study meets the criteria of a *Comparative Structured Observation* and discuss how it might be improved if it were conducted based on the recommendations in this paper.

5.1 *Comparative Structured Observation* Case Studies

Case Study 1: *Presence and Engagement in an Interactive Drama* (Dow et al., 2007)

This study is a good example of comparing two novel design variants (AR and speech based input) to an external baseline (desktop with keyboard input) in a lab setting. It also illustrates the possibility of not assigning tasks, but rather relying on an activity's inherent structure.

System: FAÇADE [20] is an interactive game that explores a couple's marital conflict. Players move through their apartment and interact primarily through conversation and by manipulating objects.

Condition comparison: The study compared two novel variants of FAÇADE — Augmented Reality (AR) and desktop with speech-based input (SP) — to a baseline desktop variant with keyboard input (KB).

Study design: The study was the first evaluation of the design concept. A key research goal of the AR variant was to understand how varying levels of immersion affect players' engagement in the game, in this case an interactive drama. The lab study was a one-factor, within-participant design that exposed participants to the three FAÇADE variants, counterbalanced for order. Recorded data included detailed logs of participants' head and body positions, as well as video of the participant's view and a third-party view of the participant. The researcher observed all participant interactions with each system and conducted an open-ended interview after game play where participants compared and reflected on their experiences with each variant. The focus was on qualitative comparison, but quantitative measures such as length of play and descriptive statistics were also reported. Results showed that the AR variant increased most participants' sense of presence with respect to the characters, the space and the story. More surprisingly, they also found that an excessive sense of presence interfered with player's engagement — some players felt too close to the action and wanted to withdraw. The authors argued that capturing and contrasting specific, measurable social cues should be included in future studies of the AR variant of FAÇADE. The authors also felt it would be valuable to test a hypothesis about social cues that emerged from the study.

Analysis as Comparative Structured Observation: The study was designed to elicit rich participant reactions to a complex experience. The authors acknowledge that game play might have been more natural had it occurred in a field setting, but given the restrictions imposed by the equipment required for the AR variant, all three conditions were conducted in the lab, which facilitated comparison. The authors did not assign specific tasks beyond playing the game itself, but took advantage of the inherent sequencing of story events as the basis for comparing

players' experiences. The authors imply implications for design from the findings rather than calling them out explicitly.

Possible improvements: If this study had been conducted as a full *Comparative Structured Observation*, the authors would have identified the specific experiences more clearly, expanded the level of comparative reflection and included explicit implications for design.

Case Study 2: *PageLinker: Integrating Contextual Bookmarks Within a Browser* (Tabard et al., 2007)

This study is a good example of comparing a new design and a baseline, and illustrates a quasi-experimental design run in the field.

System: PageLinker [51] is a context-aware bookmark designed to help research biologists navigate through and retrace chains of web-based databases and computation tools. Users can track successful transitions from one page to another, for example by embedding a contextual bookmark in a page containing source data that points directly to another with an appropriate graphing algorithm.

Condition comparison: The study compared two versions of each participant's web browser, with and without the PageLinker plugin installed.

Study design: The study was the first evaluation of the design concept. PageLinker's design was influenced by formative studies with biologists who had difficulty keeping track of previously visited web pages and the study tasks were drawn directly from this work. The longitudinal field study used an ABAB within-participants quasi-experimental design conducted over a period of four weeks: Participants alternated between weeks without PageLinker (A = baseline condition) and weeks with the PageLinker plug-in active in their personal web browser (B = experimental condition), which allowed researchers to capture data from a real-world context. Logged data included search time, clicks and page loads. Each week, the researcher observed participants perform five pre-determined tasks, counter-balanced for order when relevant. The researcher interviewed each participant at the end of the study and again after three months. In addition to reporting significant improvement in performance measures, including reduced completion time, number of clicks and pages loaded the article also reported qualitative findings about PageLinker's effect on participants' work practices over time. Findings included increased ability to return to previously discovered pages, as well as improved ability to manage interruptions and avoid "bookmark overload". The authors also identified several user innovations, including creating alternative bookmarks to avoid future breakdowns and making 'fuzzy groupings' of related pages to escape the hierarchy imposed by global bookmarks and link organisers. The qualitative findings formed the basis of the proposed implications for the design of navigation tools.

Analysis as Comparative Structured Observation: The month-long field study was designed to gather data from and about participants' experiences with PageLinker. Alternating weeks with and without PageLinker allowed participants to reflect on its effect on their daily work practices, and ensured that results were not due simply to increased proficiency over time. The imposition of a set of ecologically valid tasks let participants experience a compressed version of events that otherwise would have occurred over a longer time period. At the three-month follow-up, many participants reported that having PageLinker created a more stable, robust, navigation environment in the face of on-going change than not having PageLinker. Had the study been conducted in a lab, participants might have experienced benefits of PageLinker for the pre-defined tasks, but would have lacked insights about its impact on their daily work. By contrast, a standard field study without imposed tasks would

have made it difficult to measure longer-term performance improvements, and reduced participants' ability to make qualitative comparisons with the status quo.

Possible improvements: If this study had been conducted as a full *Comparative Structured Observation*, the authors would have highlighted the participants' comparative reflections more.

Case Study 3: *SemanticCollage: Enriching Digital Mood Board Design with Semantic Labels* (Koch et al., 2020)

This study is a good example of a two-factor experiment design where two design variants are compared in each of two design phases, resulting in a [2x2] within-participant design.

System: SemanticCollage [30] is a digital mood board that lets designers transform visual ideas into search queries. An intelligent algorithm generates semantic labels from images proposed by the designer, which results in new images and new terms that describe the images.

Condition comparison: The study compared two versions of the SemanticCollage digital mood board: WITH and WITHOUT semantic labels, as well as two mood board design phases: COLLECTION and REFLECTION.

Study design: The study was the first evaluation of the design concept. SemanticCollage's design was influenced by formative work with professional designers who had difficulty translating vague, visual ideas into productive image search queries. The study was explicitly described as a *Structured Observation*, and presented professional designers with a simulated design competition, with a compressed time frame, to generate ecologically valid tasks. Participants were exposed to a [2x2] within-participant design with four equivalent design prompts, counter-balanced across participants, followed by a fifth, open-ended condition where participants design their own mood board and freely choose whether or not to use SemanticCollage. Participants found the semantic labels particularly useful in the REFLECTION condition, and appreciated the possibility of searching for images with images. Participants reported that they felt more in control when they could see the semantic labels generated by SemanticCollage, and were better able to articulate in words what they wanted to communicate with images.

Analysis as Comparative Structured Observation: The study was designed to obtain grounded, qualitative data about participants' experiences using SemanticCollage at different phases of their design process. Participants experienced an intense, but highly realistic set of design activities, which allowed them to reflect on and compare their experiences as they tried to accomplish different goals. Researchers observed how each designer incorporated SemanticCollage into their personal design practice. Although the authors recorded and analyzed quantitative data, the primary conclusions were based on the qualitative data, analyzed with a mixed thematic analysis. The study results directly influenced the design of a follow-on system called Image Sense [30].

Possible improvements: If this study had been conducted as a full *Comparative Structured Observation*, the authors would have expanded the level of comparative reflection both within and across participants.

Case Study 4: *Designing an Eyes-Reduced Document Skimming App for Situational Impairments* (Khan et al., 2020)

This study is a good example of a new design compared to an imperfect baseline conducted in a field setting.

System: Skimmer [29] is a smartphone app that supports *eyes-reduced* auditory skimming of structured documents, such as research articles, for users who experience situational impairments, such as motion sickness when trying to read on a bus. Users can use Skimmer to skim articles while minimally using their eyes; with an easy-to-understand read-aloud overview of the document; gestures for eyes-free non-linear document navigation;

haptic cues for selective visual opt-in, such as vibrations to signal the presence of figures and tables and auditory cues that reinforce significant skimming moments, such as transitioning from one paragraph to the next.

Condition comparison: The study compared Skimmer to the VoiceDreamReader, a representative read-aloud app that served as a baseline condition.

Study design The Skimmer study was the first evaluation of the design concept. The design was influenced by a formative, needs-finding lab study that produced the concept of eyes-reduced skimming and a set of design guidelines. The study was conducted in the field, on a public city bus. The researcher sat next to the participant to ensure participant safety and observe all interactions. An audio splitter allowed the researcher to hear the same content as the participant. The study was a one-factor within-participants design that compared Skimmer to VoiceDreamReader. Participants were asked to skim two equivalent documents, isomorphic in difficulty according to several measures. Presentation of conditions and documents were counter-balanced. The article focused on qualitative comparisons, but included supplemental material with statistical analyses of additional quantitative measures, such as reading comprehension scores. The study validated the overall design concept by showing that Skimmer helped participants understand the gist of the document without using their eyes, unlike the baseline app which required heavy use of their eyes. The authors also discussed other design elements requiring further design iteration, such as the presentation of figures and tables, which users mostly ignored, even with the haptic nudge.

Analysis as Comparative Structured Observation: Participants experienced both Skimmer and VoiceDreamReader baseline app, which grounded participants' audio skimming experience. The goal was not to evaluate the two "head to head", since VoiceDreamReader was not designed for auditory skimming, but rather to assess qualitatively how the participants experienced the unique features of Skimmer in a realistic but situationally impaired setting. The authors chose a field study over a lab study to increase ecological validity, especially with respect to the participants' ability to compare real-world experiences. Half the participants reported, without prompting, that they typically experience motion sickness in a moving bus, which clearly had a strong impact on their experiences with Skimmer: participants were unable to "cheat" by reading with their eyes, as they might have in a lab study. However, the study lasted only about an hour, of which some time was "lost" waiting for the bus and other distractions. A focused hour in the lab that let participants exercise more of the design elements would have highlighted different qualitative results.

Possible improvements: If this study had been conducted as a full *Comparative Structured Observation*, the authors would have increased the number of participants to increase the richness of the qualitative insights.

Table 3 summarizes the characteristics of these four case studies with respect to *Comparative Structured Observation*, as well as additional study information to aid comparison. Note that the first two studies were published well before *Comparative Structured Observation* was characterized as an HCI method. The two latter studies were published 13 years later, after we began to conceptualize the method and were designed to follow the method's general approach, but before we had fully characterized it.

To further clarify the boundary between what what we consider to be a *Comparative Structured Observation*, we provide three "near" examples that include some structure and comparison, but do not fulfill all the criteria and thus are not considered *Comparative Structured Observation*. We also provide three examples that might on the surface appear to be *Comparative Structured Observation* but are actually not because they lack a comparison and an experimental structure. (Refer to the Appendix for four published examples from in the HCI literature that meet the criteria.)

Article Analysis		Published HCI research articles that use Comparative Structured Observation			
		Façade *Dow et al., 2007*	PageLinker *Tabard et al., 2007*	Semantic Collage *Koch et al., 2020*	Skimmer *Kahn et al., 2020*
Comparative Structured observation characteristics	Study goals	assess immersiveness types on presence & engagement	assess contextual book- marks on activity chains	assess semantic labels for composition & reflection	assess eyes-reduced design on skim & comprehend
	Design comparison	2 variants vs. baseline	new variant vs. baseline	2 variants x 2 task types [2x2]	new variant vs. baseline
	Study design	experiment	quasi experiment	experiment	experiment
	Assigned tasks	no	yes	yes	yes
	Setting	lab	field	lab	field
	Formative research activities	lab study: design variations, game forum mining	observation, participatory design workshops	participatory design workshop, questionnaire	lab study to identify design guidelines
	Participant reflection source	post-hoc interview, questionnaire	post-hoc interview, 3-month follow-up	post-hoc interview, questionnaire	post-hoc interview
	Researcher reflection source	in situ observation	in situ observation	in situ observation	in situ observation
	Analytic emphasis	qualitative, reflective	qualitative, reflective	qualitative, reflective	qualitative, reflective
	Qualitative data	video recording, researcher notes	researcher notes	video recording, researcher notes	researcher notes
	Quantitative data	logged data, questionnaire	logged data, questionnaire	questionnaire	questionnaire
Additional study details	Author description	qualitative study	time-series quasi experiment	structured observation	structured observation
	Participants	12	12	15	6
	Participant allocation	within	within	within	within
	Key comparison factor	[3] within-system variants	[2] cross-system variants (Y/N)	[2x2] variants x system or baseline	[2] system or baseline
	Structure of activity	game events	5 tasks per condition	1 task per 4 conditions	1 task per condition
	Time frame	1 3-hour session	4 1-hour sessions + followup	1 1-hour session	1 1-hour session
	Duration	single session	longitudinal	single session	single session
	Quantitative data	event logs, measures of body position, user dialogs & system processing	event logs of clicks, task completion time & page loads, questionnaire	NASA TLX questionnaire, likert-style questionnaire	reading comprehension scores
	Qualitative data	observed events, user comments	observed events, user comments	observed events, user comments	observed events, user comments
	Key outcomes	new hypothesis, implicit design implications	value of design concept, design implications, user innovations	value of design concept, design implications	value of design concept, design implications

Table 3. Comparative analysis of four published *Comparative Structured Observation* studies from the HCI research literature, according to the characteristic identified in Section 2 as well as additional study details.

5.2 Examples considered "near" *Comparative Structured Observation*

The first three examples include many of the characteristics of a *Comparative Structured Observation*, but their study designs lack at least one key element. These studies might be described as "leveraging its methodology, but are not *Comparative Structured Observation*". Note that even though a particular study design may not meet the method's criteria, this does not imply the method is not legitimate. It simply means that the researcher should not claim that the study is a *Comparative Structured Observation* and either qualify it as another established method, or justify it on its own terms.

Example 1: *Social CheatSheet: An Interactive Community-Curated Information Overlay for Web Applications.* (Vermette et al., 2017)

System: Social CheatSheet [55] overlays relevant community-curated instructions and multi-step tutorials directly atop any web application.

Condition comparison: Social CheatSheet is evaluated but not compared to any other design or baseline.

Primary evaluation goal: To assess Social CheatSheet's usefulness and usability, how well it supports the design requirements derived by the authors, and its likelihood of adoption.

Study design: The one-week "task-based field deployment" logged participant interactions and researchers probed for participants' perceptions of the system in a follow-up interview.

Why this is not Comparative Structured Observation: The comparison to the status quo exists outside of the study — participants are only asked to describe to what extent would they would like to continue to use Social Cheat-Sheet. The lack of comparison across the study conditions increases the likelihood of threats to validity. To be considered a *Structured Observation*, the researchers would have had to include direct comparisons to the status quo as an explicit part of the study, for example in an A-B-A style quasi-experimental design.

Example 2: *How Novices Sketch and Prototype Hand-Fabricated Objects.* (Bosseau et al., 2016)

System: The study does not evaluate a design concept [6] .

Condition comparison: The [3x4] mixed study design compared drawings according to two factors: audience (self, partner, external jury) and design phase (ideation, concept development, fabrication, presentation).

Primary evaluation goal: The authors are interested in designing tools to support drawing support. The study examines designers' drawings at phase of the design process for three different audiences: themselves, their partner and an external jury.

Study design: Study participants participated in a one-day phased design charette in which they produced sketches related to the design of a pair of physical artifacts. Participants created sketches for themselves, for a remote partner who would fabricate one of their designs and for an external jury.

Why this is not Comparative Structured Observation: The study design exhibits all the characteristics of the method, except that the participants are not asked to reflect on a novel design concept, nor do they compare and reflect upon their drawing strategies in each condition. Evolving a design concept is central to the method and participant comparison and reflection are needed to provide rich qualitative data that captures the participant's perspectives to mitigate against the threat of the researcher imposing their own interpretation of the observed data.

Example 3: *Feeling Stressed and Unproductive? A Field Evaluation of a Therapy-Inspired Digital Intervention for Knowledge Workers.* (Chow et al., 2023)

System: An intervention inspired by cognitive behavioral therapy that consists of: (1) using the term "Time Well Spent" in place of "productivity", (2) a mobile self-logging tool for logging activities, feelings, and thoughts at work, and (3) a visualization that guides users to reflect on their data [13].

Condition comparison: The study compared the therapy-inspired intervention to a baseline intervention that was designed to be a basic productivity-focused self-monitoring tool. Study participants only experienced one intervention, i.e., a between-participants design.

Primary evaluation goal: To assess the impact of the therapy-inspired intervention on knowledge workers compared to a classic, productivity-focused baseline intervention, and to gain insights on how to evolve the design.

Study design: Study participants participated in a four-week remote study. The duration was split into two phases: logging (while intervention was used) and follow-up (intervention not used), each lasting approximately two weeks. Surveys were administered before the logging phase (pre), after the logging phase (post), and at the end of the follow-up phase (final). Interviews were conducted at the end.

Why this is not Comparative Structured Observation: The study design exhibits all the characteristics of the method, except that the participants only experience one condition. While those who experience the therapy-inspired

intervention (novel design concept) are asked to reflect on its design, they do not compare and reflect on it relative to the baseline intervention. Thus, only the researchers can compare the experiences of the two groups of participants who experience the different interventions. The design does not mitigate against the threat of the researcher imposing their own interpretation on the observed data.

5.3 Examples considered "far" from *Comparative Structured Observation*

Many well-designed and well-conducted studies reported in the literature involve observation, include some amount of structure to the study design and use qualitative methods as the dominant analytic approach. Although these might appear at first glance to be *Comparative Structured Observation*, we do not include them because they lack an explicit comparison. We chose the following examples to help clarify the boundary between what should and not be considered *Comparative Structured Observation*.

Example 4: *Note to Self: Examining Personal Information Keeping in a Lightweight Note-Taking Tool* (Van Kleek et al., 2009)

System: LIST.IT [54] is a simple browser-based textual note-taking utility designed to capture personal information.

Condition comparison: The system is not compared to any other system.

Primary evaluation goal: The authors are interested in examining how people use personal note-taking tools and in developing a basic note-taking tool that effectively addresses their needs.

Study design: On each day of the 10-day study, participants received two email prompts to create notes. Participants completed a short exit survey at the end, but were not interviewed. Software logging captured all interactions with the system, and were analyzed with descriptive statistics. No additional qualitative data was collected.

Why this is not Comparative Structured Observation: The study design is similar to the field deployment of SOCIAL CHEATSHEET described in Example 1 in Section 5.2, but does not collect qualitative data or ask participants to compare their experiences with alternative designs. Thus the comparison and design implications are necessarily more limited than they could have been.

Example 5: *Left-over Windows Cause Window Clutter... But What Causes Left-over Windows?* (Wagner et al., 2013)

System: WM-LISA [57] is a window manager logging system developed to capture data about how users create and cope with left-over windows between work sessions.

Condition comparison: None. WM-LISA is used to log data, but is not the focus of the study.

Primary evaluation goal: To provide detailed data logs of when users create and delete windows on their desktops and how long those windows persist from session to session.

Study design: The 10-day study used WM-LISA to log when users opened and closed their windows. During the final four days, the system popped up a mini-questionnaire whenever the participant started a new session, based on five randomly chosen screenshots from their current set of left-over windows. Participants were encouraged to avoid speculation and offer specific, in-context reasons as to why they abandoned certain windows. The researcher interviewed participants at the end of the study and asked them to reflect on their reboot strategies and their patterns of left-over windows.

Why this is not Comparative Structured Observation: Participants were not asked to experience and reflect on a new system. Although participants did provide rich, qualitative data about their window management activities, they were not asked to make comparisons. The only exception was when they considered how their strategies

changed before and after a reboot, but reboots were not included directly in the study. The article focused on the quantitative analysis of participants' window management strategies, which led to implications for window management design.

Example 6: *High Costs and Small Benefits: A Field Study of How Users Experience Operating System Upgrades* (Vitale et al., 2017)

System: The study does not evaluate a new design concept [56].

Condition comparison: None

Primary evaluation goal: To understand how users experience operating system upgrades, both at the time of the upgrade and for next four weeks following the upgrade.

Study design: The field study involved in-situ observation, often in the participants' homes, where the researcher observed the participant upgrade the operating system on their own device, followed by a semi-structured interview and short survey. In the subsequent diary study, participants received a daily email reminder over the next four weeks to report any noticeable changes in their system stemming from the upgrade, or if nothing remarkable happened. The researcher conducted a brief check-in interview after two weeks and conducted a semi-structured interview at the end.

Why this is not Comparative Structured Observation: Although the study collects rich, qualitative data and participants are asked to reflect on their experiences, it does not make any explicit or implicit comparisons with a design variant.

Publication	Case study or Example	Condition comparison	Evaluation goal	Study design	CSO?	Why not *Comparative Structured Observation?*	System
Dow et al. (2007)	CS #1	2 variants vs. baseline	assess design concept	single session lab study using game events	CSO	-	FAÇADE
Tabard et al. (2007)	CS #2	new variant vs. baseline	assess design concept	limited-series field experiment, 5 representative tasks	CSO	-	PAGELINKER
Koch et al. (2020)	CS #3	2 variants x 2 task types [2x2]	assess design concept	single session lab study	CSO	-	SEMANTIC COLLAGE
Kahn et al. (2020)	CS #4	new variant vs. baseline	assess design concept	single session field study	CSO	-	SKIMMER
Vermette et al. (2017)	EX #1	-	assess design concept	1-week task-based field study	near	no condition comparison, other than informal	SOCIAL CHEATSHEET
Bosseau et al. (2016)	EX #2	no variants 2-factor design [4x3]	understand user behavior	all-day lab study, directed tasks and collaborations	near	no novel design, limited participant reflection	-
Chow et al. (2023)	EX #3	new variant vs. baseline	assess design concept	4-week logged field study with surveys, follow-up interviews	near	no condition comparison by participants	therapy-inspired intervention
Van Kleek et al. (2009)	EX #4	-	assess design concept	10-day field study with prompts	far	no condition comparison, no participant reflection	LIST.IT
Wagner et al. (2013)	EX #5	-	understand user behavior	10-day logged field study with diary study prompts, follow-up interviews	far	no novel design, no condition comparison	WM LISA (logging only)
Vitale et al. (2017)	EX #6	-	understand user experience	field study, 4-week diary study, follow-up interviews	far	no novel design, no condition comparison	-

Table 4. Comparison of four published *Comparative Structured Observation* studies with three "near" examples and three "far" examples that are not *Comparative Structured Observation*.

The four case studies and six examples above illustrate that although studies involving both observation and structured activities may appear similar, only some should be considered true *Comparative Structured Observation*. Each study's method must be analyzed carefully to ensure that it covers all the characteristics of the method.

We consider the four case studies above to be good examples of *Comparative Structured Observation* and argue that the considerable variability in their details illustrates the flexibility of the method, whereas the remaining six studies

show its boundaries. Table 4 offers a comparative analysis of these studies and summarizes why we consider each to be either an example of *Comparative Structured Observation* or a not.

6 DISCUSSION

Many HCI researchers face the problem of how to effectively characterize "mid-phase" research that occurs after initial work with users has led to a potentially interesting concept or design direction, but before it makes sense to ask precise, quantitatively testable research questions. Indeed, such questions may never make sense, especially when comparing and assessing complex, multi-faceted design concepts. In such cases, researchers are more interested in obtaining deep, grounded reflections from target users either about their experience with the new design compared to an existing system or among design variants within the new system design. As we have demonstrated, the published HCI literature already includes multiple examples of well-designed, qualitative comparative studies. However, we believe the lack of a clearly named qualitative comparison method with agreed-upon evaluation criteria has limited the number of such studies and encouraged authors to focus on quantitative results to the detriment of qualitative insights. For example, the page limits and reviewing conventions at the time of initial case studies (see Section 5) led at least some of those authors to emphasize quantitative results and limited their qualitative descriptions, even though the authors considered the latter more interesting and useful.

The goal of this article is thus to explicitly name, define and characterize *Comparative Structured Observation*. We provide clear criteria for successfully running and analyzing such studies. We first describe the origins of *Comparative Structured Observation*, and then position it within the interpretivist and positivist perspectives. We then discuss in what ways it can be considered a mixed method, followed by its benefits and limitations with respect to other HCI research methods. We explain the value of explicitly labeling and characterizing it as a method, and why it is not simply a strategy for publishing "failed" experiments. Finally, we argue that *Comparative Structured Observation* can be generalized to other disciplines beyond Human-Computer Interaction, and how "relaxed" versions may also be useful for practitioners engaged in product development.

6.1 Origins of *Comparative Structured Observation*

How did we come to define Comparative Structured Observation *as a design method?*

We came to *Comparative Structured Observation* from different backgrounds. One of us was trained as an experimental psychologist, the other a computer scientist, thus we both have strongly quantitative origins. Our respective early HCI research focused predominantly on testing interactive designs with experiments or mixed-methods approaches where quantitative methods dominated. However,, we each realized that quantitative experiments were insufficient for truly analyzing the complexity of real-world interactive systems and we both shifted to a more qualitative approach. Even so, we continued to value comparison and found that asking both users and experts to weigh the trade-offs among different design possibilities offered us deeper insights about each design concept. Like many of our HCI colleagues with similar prototype-creating, quantitative-first foundations, we suffered from the lack of accepted well-defined qualitative-first methods for evolving our designs.

We started working together nearly ten years ago during a sabbatical year when we began to co-supervise graduate students. We struggled with how to evolve the interactive designs emerging from our collaboration, spending considerable time on design and evaluation methods. This sparked deep methodological discussions and analyses of the literature that spanned subsequent reciprocal lab visits where we compared our published study designs and began to articulate the *Comparative Structured Observation* method. We now have several students who have successfully

published research with this approach. We have presented it to our respective research groups in several workshops, having developed a slide presentation to explain *Comparative Structured Observation*, which an external colleague adapted and presented to over 80 HCI researchers at a local SIG meeting. We also started using the method with external collaborators who welcomed its addition to their practice. Then the time came to characterize the method in a citable source, much like Braun & Clark [7] did with *thematic analysis*. We distributed drafts of this paper to collaborators and external colleagues with both quantitative and quantitative backgrounds and incorporated their feedback as we iterated on the characteristics and created more precise definitions. This feedback also led us to develop the two checklists in Section 4 to help researchers adopt *Comparative Structured Observation*. This paper is the culmination of our collaboration, which we hope will help HCI researchers adopt *Comparative Structured Observation* into their research practice.

6.2 Epistemological underpinnings

Where does Comparative Structured Observation fit with respect to interpretivist vs. positivist perspectives?

Orlikowski & Baroudi [44] explain that "positivist" studies are designed to test theory in order to increase predictive understanding of phenomena, and rely upon quantifiable measure of variables, hypothesis testing, and drawing inferences to generalize from a sample to a population. By contrast, they note that "interpretivist" studies assume that people create subjective meaning as they interact in the world. The researchers who conduct such studies reject the idea of an "objective" or factual account of events or situations, and, rather than generalizing from the specifics, seek to understand the phenomenon more deeply within its context. They argue that *"researchers should ensure that they adopt a perspective that is compatible with their own research interests and predispositions, while remaining open to the possibility of other assumptions and interests."*

Structured Observation aligns strongly with an interpretivist perspective, but with an infusion of positivism. *Structured Observation* is not appropriate for testing formal theory, nor uncovering any objective truth, nor claiming causality or predictive power. Thus, it is not staunchly positivist. However informal hypotheses or "hunches" about how elements in a design concept might be experienced can be investigated systematically with *Structured Observation* by leveraging experimental methods. The term "hypothesis" has a very specific meaning in experiment design, where a system can be viewed as "better" based on quantitative metrics such as improved performance. By contrast, informal hunches about qualitative differences between design variants do not lead easily to claims about which is "better", since they cannot be directly quantified. Hypotheses are often not possible due to lack of dependent variables or present dependent variables that are not amendable to statistics due to insufficient control of conditions and low numbers of participants. Experimental methods can be leveraged in *Structured Observation*, in particular, by assigning tasks and structuring participants' exposure to design variants to increase rigor. Adding structure to the imposition of tasks should result in more robust findings from the qualitative data. For example, exposing a participant to a single design variant is more prone to a participant (even unknowingly) wanting to please the researchers, perhaps by being overly positive about the design variant, than if they compare two or more variants. Comparison, and in particular reflecting on experiences with more than one design variant, naturally involves critique, opening the door for participants to make both positive and negative remarks.

In what ways can Comparative Structured Observation be considered a mixed-method approach?

We define *Structured Observation* as a qualitative method to emphasize that it can be used solely with the collection and analysis of qualitative data. The method is also interventionist, in that researchers control the presentation of tasks to participants. Finally, it can be considered a mixed-methods approach in two senses. First, it always involves

two qualitative methods: observation and interviewing. Although HCI researchers have often combined methods as a way to triangulate data and solidify their findings, the traditional definition of "mixed methods" has been to combine quantitative and qualitative methods in a single study, where quantitative methods typically dominate and qualitative methods are secondary. HCI researchers often use qualitative analyses to help explain quantitative findings. *Structured Observation* inverts these priorities by emphasizing qualitative data collection and analysis and making quantitative analysis secondary or even optional. Even so, a *Comparative Structured Observation* study may sometimes benefit from including performance or other quantitative measures. For example, if users react positively to a novel system according to various subjective criteria, the designer may still want to ensure that performance is as good, or at least not "worse". Capturing quantitative measures can also highlight any mismatches between the user's perception of their performance and their actual performance. For example, users may not realize how much time has actually elapsed as they perform a particularly interesting (or annoying) task, which may have a corresponding impact on future design decisions.

The second way *Structured Observation* can be considered a mixed-methods approach is that quantitative methods always involve multiple components. Usually an experiment protocol is designed, quantitative data is collected and that data is analyzed using quantitative analysis techniques. Similarly qualitative methods usually involve designing an instrument such as an interview protocol, collecting qualitative data and then analyzing that data. Because *Structured Observation* does not reject the benefits of using an experiment protocol it can thus also be viewed as a mixed-method approach that combines elements of both quantitative and qualitative methods.

6.3 Benefits of *Comparative Structured Observation*

We summarize the benefits of *Comparative Structured Observation* for HCI researchers who seek a systematic qualitative method for working with users to *"get the right design"*. The method offers multiple benefits beyond those inherent in observing and interviewing users who interact with design artifacts, including:

- providing a systematic method for assessing both low- and medium-fidelity prototypes allows researchers to more rigorously assess design concepts at an earlier phase of a design project.
- comparing ecologically valid interactions with multiple design variants encourages users to critique rather than simply accept design variants and contribute both positive and negative comments.
- basing comparisons on the details of users' recent, lived experience as they perform the tasks takes advantage of human memory to produce better, more grounded results.
- comparing users' reflections on their own experience to researcher's observations of the same experience provides additional insights, since users' perceptions and actions contribute differently to understanding the design.
- generating sources of comparison both from participants and researchers encourages researchers to reconcile situations where users' reflections about their experiences differ from actions observed by the researcher, which can be used to productively advance the design.
- structuring users' activities into controlled experiences (tasks and conditions) can reduce analysis time relative to more open-ended study designs and their associated analysis, e.g. by focusing on bottom-up inductive thematic analysis.

6.4 Value of labeling and characterizing *Comparative Structured Observation*

Our contribution lies in labeling and richly characterizing *Comparative Structured Observation*, with detailed criteria for determining whether a particular study should be considered as a *Structured Observation* and quality criteria for assessing whether or not a particular study is considered "good". We believe that naming and characterizing *Structured Observation* is valuable for two key reasons. First, it provides greater clarity about the research that *we already do*. So, even if it does not change our practice, it will enable a crisper and more standardized way of articulating a study's method. This has a number of side benefits such as supporting novice researchers as they learn about this method, as well as benefits for readers of articles or manuscripts in which a *Structured Observation* was conducted. Using a named method in a manuscript will help to set readers' expectations and should help standardize peer review.

Second, naming and characterizing the method *may change what we do*. Although some examples of *Structured Observation* exist in the literature, both named and not named, they are relatively rare compared to non-comparative observational field studies or controlled experiments. We hope that naming and characterizing *Structured Observation* will lend credibility to this method, thereby giving permission to researchers to use it, perhaps at times when they might have previously felt the need to "sandwich" their work into an experiment or usability test [25]. Historically, experiments served as the gold standard in HCI for evaluation studies, given that they are rigorous and scientific. However, experiments offer limited information for discovering the right design [25]. We hope that legitimizing *Structured Observation* will lead to the publication of more studies that probe deeply into novel design concepts in the middle phase of the iterative design cycle. This in turn may result in greater adoption of the design innovations stemming from published HCI research than we have seen to date.

Note that *Structured Observation* should not be considered as a strategy for publishing "failed" experiments. We have been asked whether researchers who have run an experiment that did not reach statistical significance could retrospectively re-frame their study as *Structured Observation* in order to make it acceptable for publication, essentially by re-targeting their findings to focus on the qualitative. This is clearly not our intention. If a testable hypothesis is within reach, e.g., that design A will be faster or less error prone than design B, then a classic experiment with quantitative results should be expected by reviewers, likely supplemented by qualitative findings. In such cases, reviewers should be wary of studies that present *only* qualitative data.

6.5 Challenges and limitations of *Comparative Structured Observation*

Like all research methods, *Comparative Structured Observation* is not a panacea. *Structured Observation* is explicitly not recommended for the earliest phases of a research project, because defining relevant tasks or user experiences often requires initial formative studies and a clearly defined design concept or direction. Although researchers can benefit from leveraging studies published in the HCI research literature to inform their design concept and tasks, we believe that *Structured Observation* is most successful when researchers perform preliminary studies of their own with users for this purpose. When researchers seek to validate a refined system's design and testable hypotheses are within reach, they should defer to controlled experiments.

As with any research method, *Structured Observation* poses certain execution challenges. Setting up an intervention that provides users with realistic, in-context experiences with multiple design variants may be difficult, especially when a novel technology that demonstrates the design concept is not sufficiently mature. Finding appropriate comparative conditions, baseline or otherwise, or identifying user tasks that fit within the constraints of a study can also prove difficult. Finally, some participants may find it challenging to project themselves into the tasks or activities designed

by the researcher, which may limit the value of their reflections. As always when using a qualitative interpretivist approach, the findings will never be fully replicable. *Comparative Structured Observation* cannot be used to establish causal relationships nor to make certain types of generalizations.

A number of these challenges overlap with the challenges identified by Olsen [43] when evaluating *systems research*, which emphasizes engineering, building and deploying novel interactive systems. He argues that testable hypotheses are are rarely possible, making controlled experiments a poor fit. *Structured Observation* faces a similar problem, since establishing comparative conditions can be difficult or even impossible: Novel prototype systems never match the fidelity of a mature existing baseline system nor is it feasible to fully engineer and deploy more than one design variant unless each is relatively small. *Structured Observation* does not therefore address the evaluation challenges posed by HCI-related systems research.

6.6 Generalizability to other disciplines and to practitioners

We have provided a detailed characterization of *Comparative Structured Observation* to help HCI researchers interested in advancing a novel design concept. However, the rigorous approach to comparison and reflection can also be applicable beyond HCI, for example, asking study participants to compare and reflect on non-technological experiences. Most of the methods used by HCI researchers have been appropriated from other fields that are not specific to HCI. While some methods such as Contextual Inquiry originated within HCI practice, they are the minority. Just as most other methods used by HCI researchers can be generalized beyond our use, so too should *Structured Observation*.

For example, an educational researcher might design a new lecture delivery method and decide to compare it to a status quo delivery method. They might try each method at two different points within the course and collect rich qualitative feedback from the students as well as their own observations. The researcher could assess the strengths and weaknesses of each method and use the results to iterate on the lecture delivery method, after which they could run a controlled experiment that compares quantitative learning outcomes on pre- and post-tests with a sufficiently large number of students to achieve statistical significance.

Although we explicitly target HCI researchers in this article, we also see a role for *Structured Observation* for practitioners, based on our own experiences using *Structured Observation* variations for product development. Product 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" [34, 35]. However, even a "relaxed *Structured Observation*" will let participants experience and reflect on design variants, thus providing designers insights about complex new design concepts. In this way, *Structured Observation* may offer benefits to researchers and practitioners alike.

7 CONCLUSION

This article argues that HCI researchers need a well-defined, qualitative empirical research method for obtaining deeper insights from users about novel design concepts. We present *Comparative Structured Observation*, a rigorous qualitative method that captures rich, qualitative data from study participants as they compare their experiences with different design variants. Researchers select tasks or structure experiences following well-established experimental design principles, so that study participants and researchers can compare and reflect on the design variants.

The approach can appeal to both qualitatively trained and quantitatively trained HCI researchers. Qualitative researchers may retain an interpretivist stance with respect to gathering and analyzing rich, comparative, qualitative

data, and also benefit from systematic reflection from both participants and researchers to gain new insights about complex design concepts. Quantitative researchers may leverage a more positivist stance, and benefit concretely from analyzing qualitative data about comparable tasks or user activities.

Although the focus of *Structured Observation* is to generate deeper, qualitative insights about a proposed design direction or to inspire a new one, another possible outcome is to transform informal hunches into testable hypotheses. As Louis Pasteur famously observed: "In the fields of observation, chance favors only the prepared mind."[53] Although clearly irrelevant for some research questions and never a required outcome, *Structured Observation* does offer a systematic method for "preparing one's mind" to generate new hypotheses about future designs. Finally, our own experience with *Structured Observation* leads us to believe that practitioners in industry may also benefit, although probably with somewhat relaxed criteria. Real-world design projects must balance a large number of complex factors and criteria that cannot be addressed with controlled experiments, not only because of the resources required, but because many issues lack clear performance metrics and involve multiple confounding variables. *Structured Observation* offers a method for obtaining grounded, comparative feedback from users about multi-faceted design concepts that would be difficult to assess by other means.

In summary, this article contributes *Comparative Structured Observation* as an HCI research method and specifies its characteristics. We arrived at the name and characteristics through iterative discussions and detailed consideration of our own work and other similar work. We provide a detailed checklist that researchers and reviewers can use to determine whether a study is a *Structured Observation* and elaborate on the qualities needed for a study to be considered a "good" *Structured Observation*. Our analysis of four published case studies illustrates both how they meet the criteria for *Structured Observation*, but could also be improved by following the recommendations outlined in this article. In addition, our analysis of six "near" and "far" examples helps clarify the boundary between what should and should not be considered *Structured Observation* as a step towards maturing HCI as a discipline.

ACKNOWLEDGMENTS

Our thanks to everyone who contributed to the discussions about *Comparative Structured Observation* in our respective research labs. A special thanks to Chat Wacharamanotham, Alexander Eiselmayer, and Michel Beaudouin-Lafon for their deep engagement in this project. Additional thanks go to Andrea Bunt, Parmit Chilana, Steven Dow, Jérémie Garcia, Aurélien Tabard, Mike Massimi, and Dongwook Yoon, Teerapaun (Mui) Tanprasert for their early reviews. We are also very grateful to our anonymous reviewers whose feedback significantly strengthened our manuscript.

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".

REFERENCES

[1] Lorans Alabood, Zahra Aminolroaya, Dianna Yim, Omar Addam, and Frank Maurer. 2023. A systematic literature review of the Design Critique method. *Information and Software Technology* 153 (2023), 107081. https://doi.org/10.1016/j.infsof.2022.107081

[2] M.V. Angrosino. 2007. *Naturalistic Observation (1st ed.).* Routledge, New York. https://doi.org/10.4324/9781315423616

[3] Michel Beaudouin-Lafon, Susanne Bødker, and Wendy E. Mackay. 2021. Generative Theories of Interaction. *ACM Trans. Comput.-Hum. Interact.* 28, 6, Article 45 (nov 2021), 54 pages. https://doi.org/10.1145/3468505

[4] Jekaterina Belakova and Wendy E. Mackay. 2021. SonAmi: A Tangible Creativity Support Tool for Productive Procrastination. In *C&C '21 - 13th ACM Conference on Creativity & Cognition.* ACM, Virtual Event, Italy, 1–10. https://doi.org/10.1145/3450741.3465250

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

[6] Adrien Bousseau, Theophanis Tsandilas, Lora Oehlberg, and Wendy E. Mackay. 2016. How Novices Sketch and Prototype Hand-Fabricated Objects. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 397–408. https://doi.org/10.1145/2858036.2858159

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[8] 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

[9] Andrea Bunt, Joanna McGrenere, and Cristina Conati. 2007. Understanding the Utility of Rationale in a Mixed-Initiative System for GUI Customization. In *Proceedings of the 11th International Conference on User Modeling* (Corfu, Greece) *(UM '07)*. Springer-Verlag, Berlin, Heidelberg, 147–156. https://doi.org/10.1007/978-3-540-73078-1_18

[10] Susanne Bødker. 1991. *Through the Interface: A Human Activity Approach to User Interface Design.* L. Erlbaum Associates Inc., USA.

[11] J. Carthey. 2003. The role of structured observational research in health care. *Quality and Safety Health Care* 1, 1 (2003), 13–16. https://doi.org/10.1136/qhc.12.suppl_2.ii13

[12] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis.* Sage Publications, Thousand Oaks, California.

[13] Kevin Chow, Thomas Fritz, Liisa Holsti, Skye Barbic, and Joanna McGrenere. 2023. Feeling Stressed and Unproductive? A Field Evaluation of a Therapy-Inspired Digital Intervention for Knowledge Workers. *ACM Trans. Comput.-Hum. Interact.* 31, 1, Article 12 (Nov. 2023), 33 pages. https://doi.org/10.1145/3609330

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

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

[16] Paul Croll. 1986. *Systematic Classroom Observation.* Social research and educational studies, Vol. 3. Taylor & Francis Group, New York. 202 pages.

[17] Paul Croll. 2004. Structured Observation. In *Encyclopedia of Research Methods for the Social Sciences*, A. Bryman M. Lewis Beck and T. Futing Liao (Eds.). SAGE Publications, New York, 1096–1098.

[18] Scott Davidoff, Min Kyung Lee, Anind K. Dey, and John Zimmerman. 2007. Rapidly Exploring Application Design Through Speed Dating. In *UbiComp 2007: Ubiquitous Computing*, John Krumm, Gregory D. Abowd, Aruna Seneviratne, and Thomas Strang (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 429–446.

[19] Norman K. Denzin and Yvonna S. Lincoln. 2017. *SAGE Handbook of Qualitative Research, Fifth Edition.* Sage Publications, Thousand Oaks, California. 992 pages.

[20] Steven Dow, Manish Mehta, Ellie Harmon, Blair MacIntyre, and Michael Mateas. 2007. Presence and Engagement in an Interactive Drama. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1475–1484. https://doi.org/10.1145/1240624.1240847

[21] Jérémie Garcia, Theophanis Tsandilas, Carlos Agon, and Wendy Mackay. 2012. Interactive Paper Substrates to Support Musical Creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. ACM, New York, NY, USA, 1825–1828. https://doi.org/10.1145/2207676.2208316

[22] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: Cultural Probes. *Interactions* 6, 1 (jan 1999), 21–29. https://doi.org/10.1145/291224.291235

[23] Wayne D. Gray and Marilyn C. Salzman. 1998. Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Human-Computer Interaction* 13, 3 (1998), 203–261. https://doi.org/10.1207/s15327051hci1303_2

[24] J. Greenbaum and M. (Eds.) Kyng. 1991. *Design at Work: Cooperative Design of Computer Systems (1st ed.).* CRC Press, Boca Raton, Florida, USA. 306 pages.

[25] 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

[26] Kristina Höök and Jonas Löwgren. 2012. Strong Concepts: Intermediate-level Knowledge in Interaction Design Research. *ACM Trans. Comput.-Hum. Interact.* 19, 3, Article 23 (Oct. 2012), 18 pages. https://doi.org/10.1145/2362364.2362371

[27] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology Probes: Inspiring Design for and with Families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. ACM, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[28] R. Johnson and Anthony Onwuegbuzie. 2004. Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational researcher* 33 (10 2004), 14. https://doi.org/10.3102/0013189X033007014

[29] Taslim Arefin Khan, Dongwook Yoon, and Joanna McGrenere. 2020. *Designing an Eyes-Reduced Document Skimming App for Situational Impairments.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376641

[30] Janin Koch, Nicolas Taffin, Andrés Lucero, and Wendy E. Mackay. 2020. SemanticCollage: Enriching Digital Mood Board Design with Semantic Labels. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. ACM, New York, NY, USA, 407–418. https://doi.org/10.1145/3357236.3395494

[31] Karl Lewin. 1947. Frontiers in Group Dynamics. *Human Relations* 1 (1947), 5–41. https://doi.org/10.1177/001872674700100103

[32] Henry Lieberman. 2003. *The Tyranny of Evaluation.* ACM, New York, NY, USA. https://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html

[33] Csikszentmihalyi M and Larson R. 1987. Validity and reliability of the Experience-Sampling Method. *J Nerv Ment Dis.* 175, 9 (Sep 1987), 526–36. https://doi.org/10.1097/00005053-198709000-00004

[34] Wendy E. Mackay. 2019. Designing with Sticky Notes. In *Sticky Creativity: Post-It Note Cognition, Interaction and Digitalization*, Bo Christensen, Kim Halskov, and Clemens Klokmose (Eds.). Academic Press, Cambridge, MA, USA, 231–256. https://hal.science/hal-02426827

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

[36] Wendy E. Mackay and Anne-Laure Fayard. 1997. HCI, Natural Science and Design: A Framework for Triangulation Across Disciplines. In *Proceedings of the 2nd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (Amsterdam, The Netherlands) *(DIS '97)*. ACM, New York, NY, USA, 223–234. https://doi.org/10.1145/263552.263612

[37] W. E. Mackay and Anne-Laure Fayard. 1999. Designing Interactive Paper: Lessons from Three Augmented Reality Projects. In *Proceedings of the International Workshop on Augmented Reality : Placing Artificial Objects in Real Scenes: Placing Artificial Objects in Real Scenes* (Bellevue, Washington, USA) *(IWAR '98)*. A. K. Peters, Ltd., Natick, MA, USA, 81–90. http://dl.acm.org/citation.cfm?id=322690.322698

[38] Mark J. Martinko and William L. Gardner. 1990. STRUCTURED OBSERVATION OF MANAGERIAL WORK: A REPLICATION AND SYNTHESIS*. *Journal of Management Studies* 27, 3 (1990), 329–357. https://doi.org/10.1111/j.1467-6486.1990.tb00250.x

[39] Scott E Maxwell, Harold D Delaney, and Ken Kelley. 2017. *Designing experiments and analyzing data: A model comparison perspective.* Routledge, Taylor & Francis Group, Abingdon-on-Thames, England, UK. 1080 pages.

[40] Joseph E. McGrath. 1995. Methodology Matters: Doing Research in the Behavioral and Social Sciences. In *Readings in Human–Computer Interaction*, Ronald M. BAECKER, Jonathanb GRUDIN, William A.S. BUXTON, and Saul GREENBERG (Eds.). Morgan Kaufmann, New York, 152–169. https://doi.org/10.1016/B978-0-08-051574-8.50019-4

[41] Henry Mintzberg. 1970. Structured observation as a method to study managerial work. *Journal of Management Studies* 7, 1 (1970), 87–104. https://doi.org/10.1111/j.1467-6486.1970.tb00484.x

[42] Jakob Nielsen. 1994. Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies* 41, 3 (1994), 385–397. https://doi.org/10.1006/ijhc.1994.1065

[43] Dan R. Olsen. 2007. Evaluating User Interface Systems Research. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology* (Newport, Rhode Island, USA) *(UIST '07)*. Association for Computing Machinery, New York, NY, USA, 251–258. https://doi.org/10.1145/1294211.1294256

[44] Wanda J. Orlikowski and Jack J. Baroudi. 1991. Studying Information Technology in Organizations: Research Approaches and Assumptions. *Information Systems Research* 2, 1 (1991), 1–28.

[45] Antti Oulasvirta and Kasper Hornbæk. 2022. Counterfactual Thinking: What Theories Do in Design. *International Journal of Human–Computer Interaction* 38, 1 (2022), 78–92. https://doi.org/10.1080/10447318.2021.1925436

[46] John Rieman. 1993. The Diary Study: A Workplace-Oriented Research Tool to Guide Laboratory Efforts. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 321–326. https://doi.org/10.1145/169059.169255

[47] P.J. Runkel and J.E. McGrath. 1972. *Research on Human Behavior: A Systematic Guide to Method.* Holt, Rinehart and Winston, New York, NY, USA. https://books.google.dk/books?id=3FdqAAAAMAAJ

[48] W. R. Shadish, T. D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2 ed.). Houghton Mifflin, New York.

[49] Erik Stolterman and Mikael Wiberg. 2010. Concept-Driven Interaction Design Research. *Human–Computer Interaction* 25, 2 (2010), 95–118. https://doi.org/10.1080/07370020903586696

[50] 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.

[51] Aurélien Tabard, Wendy Mackay, Wendy Mackay, Nicolas Roussel, and Catherine Letondal. 2007. PageLinker: Integrating Contextual Bookmarks Within a Browser. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. ACM, New York, NY, USA, 337–346. https://doi.org/10.1145/1240624.1240680

[52] Kimberly Tee, Karyn Moffatt, Leah Findlater, Eve MacGregor, Joanna McGrenere, Barbara Purves, and Sidney S. Fels. 2005. A Visual Recipe Book for Persons with Language Impairments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 501–510. https://doi.org/10.1145/1054972.1055042

[53] René Vallery-Radot. 1902. *The Life of Pasteurs.* McClure, Phillips and Company, New York, NY, USA.

[54] Max G. Van Kleek, Michael Bernstein, Katrina Panovich, Gregory G. Vargas, David R. Karger, and MC Schraefel. 2009. Note to Self: Examining Personal Information Keeping in a Lightweight Note-Taking Tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1477–1480. https://doi.org/10.1145/1518701.1518924

[55] Laton Vermette, Shruti Dembla, April Y. Wang, Joanna McGrenere, and Parmit K. Chilana. 2017. Social CheatSheet: An Interactive Community-Curated Information Overlay for Web Applications. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 102 (dec 2017), 19 pages. https://doi.org/10.1145/3134737

[56] Francesco Vitale, Joanna McGrenere, Aurélien Tabard, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2017. High Costs and Small Benefits: A Field Study of How Users Experience Operating System Upgrades. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*

(Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4242–4253. https://doi.org/10.1145/3025453.3025509

[57] Julie Wagner, Mathieu Nancel, Sean Gustafson, Stéphane Huot, and Wendy E. Mackay. 2013. A Body-centric Design Space for Multi-surface Interaction. In *CHI'13 - 31st International Conference on Human Factors in Computing Systems*. ACM SIGCHI, ACM, Paris, France, 47-50. https://hal.inria.fr/hal-00789169

8 APPENDIX

8.1 Characteristics of a well-constructed *Comparative Structured Observation* study

Table 5 summarizes the key characteristics of *Comparative Structured Observation* as well as optional characteristics. Characteristics that are considered incorrect are listed in the right-hand column. (Refer to Table 2 in Section 4 for a more detailed checklist of what to include.)

Required CSO characteristics	Optional CSO characteristics	Incorrect CSO characteristics
Researchers explicitly define comparable design variants.	Lab and field settings are both appropriate.	Compares design to status quo only, outside the context of the study.
Researchers derive participant tasks based on formative research.	Participants use a talk aloud protocol.	Runs an open-ended field test with no comparisons.
Participant exposure to tasks and design variants meets experimental design best practices.	Quantitative data informs qualitative data.	Omits comparable tasks or experiences.
Participants experience at least two design variants within the study.	Researchers generate testable hypotheses.	Omits participant reflection on comparable experiences.
Participants compare and reflect upon their experiences with each design variant.		Omits researcher observation and reflection on participants' experiences.
Researchers observe participants' experiences directly or through other rich data sources.		Focuses only on performance metrics.
Researchers emphasize gathering qualitative data.		
Researchers conduct post-hoc interviews.		

Table 5. A correct *Comparative Structured Observation* includes all required characteristics listed in column 1, and may include optional characteristics listed in column 2. The characteristics listed in column 3 do not belong in a *Comparative Structured Observation*.

8.2 Additional published examples of *Comparative Structured Observation*

The following examples illustrate four additional variations of *Comparative Structured Observation* published in the HCI literature.

Example 7: *Interactive Paper Substrates to Support Musical Creation.* (Garcia et al., 2012)

System: Polyphony [21] is an interactive paper-based composition tool that lets composers design their own musical structures.

Condition comparison: The [2x2] within-participants design compares Polyphony with baseline music composition tools on two tasks: original composition and modification of an existing composition.

Primary evaluation goal: To assess the design concept: "to identify common patterns that emerge, despite the highly individual nature of composition strategies, [to] support fluid transitions between pen-based and existing software composition tools."

Study design: The lab study asked professional composers to create an original composition and enhance an existing composition in one hour. Researchers observed participants and asked them to reflect on their experiences in each condition. Note that this is the first published HCI article that defines *Comparative Structured Observation* as part of the contribution.

Example 8: *SonAmi: A Tangible Creativity Support Tool for Productive Procrastination* (Belakova & Mackay, 2021)

System: SonAmi [4] is an an interactive coaster for professional writers to make procrastination productive by speaking dialog out loud whenever the author lifts their mug.

Condition comparison: The one-factor within-participants design compares two variations of SonAmi: either a computer-generated voice or a recording of the authors' own voice.

Primary evaluation goal: To assess and compare two design variants: "to compare human-generated and synthetic voices". A separate field study used SonAmi as a technology probe.

Study design: The two-day field study asked professional first asked writers to record their own voices speaking dialog they had written. The same dialog was also recorded using a computer-generated voice. After a briefing session, participants were asked to perform two equivalent writing tasks over a period of two days, one using the "Own Voice" condition and the other using the "Computer Voice" condition. The researcher conducted a final interview to compare the two and reflect on their experiences. This paper explicitly refers to using as the design method.

Example 9: *Understanding the Utility of Rationale in a Mixed-Initiative System for GUI Customization* (Bunt et al., 2007)

System : The Mica (Mixed-Initiative Customization Assistance) [9] system provides the rationale for system-suggested GUI personalizations in a feature-rich word-processing interface.

Condition comparison: The one-factor within-participants design compares the same word-processing interface with and without Mica on a series of tasks.

Primary evaluation goal: To assess the design concept: "to better understand the qualitative impact of the rationale on users' attitudes toward the system."

Study design: The lab study consisted of three-hour sessions where participants completed a "guided task". Researchers observed participants and interviewed them about their experiences under both conditions.

Example 10: *A visual recipe book for persons with language impairments.* (Tee et al., 2005)

System: VERA (Visually Enhanced Recipe Application) [52] is a multi-modal recipe application designed to support people with aphasia, an acquired language impairment.

Condition comparison: The one-factor within-participants design compared VERA to a text-based recipe that was adapted to be as aphasia-friendly as possible using status-quo techniques.

Primary evaluation goal: To assess the design concept: "Although the evaluation followed a structured experimental design so that we could have identified statistically significant trends had they arisen, we were motivated by the qualitative observations case study analyses could provide."

Study design: The field study asked participants with aphasia to complete two different recipes, counter-balanced for order, in their own kitchens. Researchers observed and interviewed participants about their experiences.