AIS Transactions on Human-Computer Interaction

Volume 9 | Issue 2 Article 4

6-30-2017

ECOVAL: Ecological Validity of Cues and Representative Design in User Experience Evaluations

Suzanne Kieffer
Université catholique de Louvain, suzanne.kieffer@uclouvain.be

Recommended Citation

Kieffer, S. (2017). ECOVAL: Ecological Validity of Cues and Representative Design in User Experience Evaluations. *AIS Transactions on Human-Computer Interaction, 9*(2), 149-172. Retrieved from https://aisel.aisnet.org/thci/vol9/iss2/4
DOI:

Transactions on Human – Computer Interaction

Research Paper

ECOVAL: Ecological Validity of Cues and Representative Design in User Experience Evaluations

Suzanne Kieffer
Université catholique de Louvain
suzanne.kieffer@uclouvain.be

Abstract:

Egon Brunswik coined and defined the concepts of ecological validity and representative design, which are both essential to achieve external validity. However, research in HCI has inconsistently and incorrectly used Brunswik's concept of ecological validity, which prevents the field from developing cumulative science and from generalizing the findings of user experience (UX) evaluations. In this paper, I present ECOVAL, a framework I built on Brunswik's ideas. On the one hand, ECOVAL helps HCI researchers describe and assess the ecological validity of cues in UX evaluations. On the other hand, ECOVAL guidelines—formulated as a step-by-step procedure—help HCI researchers achieve representative design and, therefore, increase external validity. An industrial case study demonstrates the relevance of ECOVAL for achieving representative design while conducting formative UX testing. In discussing the case study, I describe how ECOVAL can help HCI researchers assess and increase the validity of UX experiments and generalize UX findings. I also illustrate the trade-offs between internal and external validities and UX resources that inevitably arise when one conducts UX experiments. From the results, I sketch avenues for future research and discuss the related challenges that future work should address.

Keywords: Industry, Method, Experiment, Case Study.

A prior version of the paper was presented at HICSS-48, January 5-8, 2015.

The manuscript was received 05/08/2015 and was with the author for 8 months for 2 revisions.

1 Introduction

Egon Brunswik (1940, 1952) coined and defined the concept of ecological validity (EV) as the statistical correlation between a proximal cue and the distal variable to which it relates. Brunswik's double convex lens (Figure 1) shows a collection of proximal cues {C1, …, CN} that diverge from the actual state of the environment, the distal variable. Organisms use the cues to predict the distal variable, and, therefore, the cues converge at the point of inference (judgment, decision, or response) in the organism. The cues are interrelated, which introduces redundancy in the environment and, therefore, provides organisms with multiple alternatives to achieve the distal criterion (Dhami, Hertwig, & Hoffrage, 2004). Through experience, organisms learn the EV of cues, their inter-correlations, and how to cope with the environment's uncertainty.

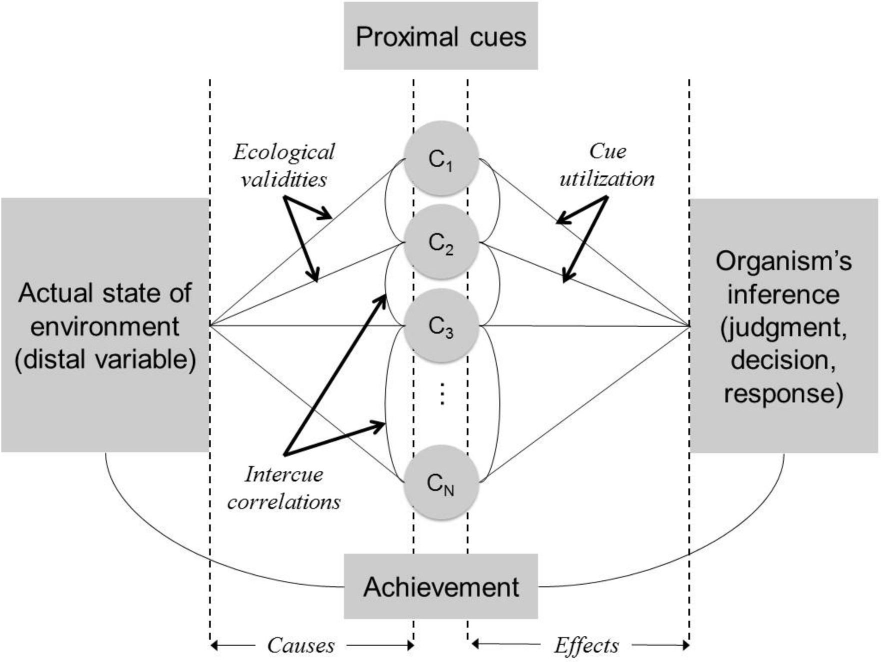


Figure 1. Brunswick (1956) Lens Model

For example, an organism can use the size of a cave aperture (proximal cue) to know whether the cave is a safe place to rest (distal variable), which it might measure as the ratio of unsafe to safe events that occur in that cave[1]. The size has an EV close to 1 because a cave's safety is highly correlated with its size. By contrast, the presence of trees' surrounding the cave has an EV close to 0 if they are likely to surround both safe and unsafe caves.

Consequently, HCI researchers must use EV to compare the quality of different cues and to understand why judgment based on such cues must have limited accuracy. Furthermore, they must use EV to assess the representativeness of an experimental design (i.e., whether the situation captured in the experimental setting enables the organism "to perceive in order to act, but also to act in order to perceive" (Araujo et al., 2007, p. 12). If they do not use EV to do so, the organism's experimental behavior may not correspond to the functional behavior toward which the researcher wishes to generalize. In their study of the relationship between simulators and simulated systems, Stoffregen, Bardy, Smart, and Pagulayan (2003) refer to this reproduction of behavior as "action fidelity". Action fidelity is what neo-Brunswikians name "task representativeness" (Dhami et al., 2004).

EV is important to anyone concerned with external validity (i.e., generalizing findings from an experiment to particular persons, times, and places) (Gray & Salzman, 1998). Brunswik (1956) argue that the number of cues, their ecological validities, and their inter-correlations should be representative of the environment

[1] I extract this example from Araujo, Davids, and Kassos (2007, p. 4).

toward which one applies experimental outcomes; otherwise, one would limit their findings' generalizability (Hammond, 1998a; Dhami et al., 2004; Araujo et al., 2007). Brunswik (1957) further points out that experimental research should equally emphasize the organism (participant sampling) and the environment; otherwise, one would make generalizations to the environment without scientific evidence (Hammond, 1998a; Dhami et al., 2004; Araujo et al., 2007). EV actually goes beyond external validity because it offers a formal way to measure the representativeness of experimental designs and the ecological relevance of the experimental task and, therefore, enables one to generalize their findings to both organisms and environments.

In human-computer interaction (HCI), EV is crucial for designing relevant and meaningful UX because UX evaluation (UXE) is a key milestone toward achieving successful UX. First and foremost, UXEs offers insights into users and their experience with a product. Decision makers rely on this information to structure and optimize the processes involved in the product's development (Tullis & Albert, 2013). Second, UXE findings directly feed into UX design (Mayhew, 1999). To not representatively sample on the environmental side (e.g., sensory stimuli, everyday objects or social interactions) may fail to capture relevant aspects of the real world and, therefore, fail to engage participants in performing the experimental task as they would have for real. As a result, uncertainty about 1) the generalization of the UXE findings, 2) the relevance of the features embedded in the product, and 3) the quality of the UX can arise.

2 Contribution

In HCI, research has inconsistently and incorrectly used EV. It has used EV to refer to fieldwork (Carter, Mankoff, Klemmer, & Matthews, 2008; Bernstein, Ackerman, Chi, & Miller, 2011), real-life or naturalistic conditions where experimental tasks are performed with high levels of fidelity (Castro, Favela, & Garcia-Pena, 2011), and studies "where subjects are totally unaware of being tested, testing being performed during their natural activity on the web" (Guerini, Strapparava, & Stock, 2012). None of these interpretations actually refer to EV. Instead, they refer to the notion of representative design, another concept that Brunswik (1944) coined. Representative design involves randomly sampling stimuli from the environment to create an experimental setting that is "representative" of the population of stimuli to which the researcher wishes to generalize (Brunswik, 1944). These inconsistencies prevent the development of cumulative science (Hammond, 1998a; Hammond, 1998b). They also make EV a confusing and misleading concept in HCI.

I argue that Brunswik's (1944) concepts of EV and representative design can help HCI researchers to frame and justify the generalization of UX findings (i.e., external validity). Furthermore, I argue that UX evaluations must report on the sampling of both the population (sample size and participants profile) and the cues to allow relevant and constructive discussion about external validity. Especially, they must report on the number of cues, their ecological validity, and their inter-correlations to provide sufficient evidence of task representativeness.

In this paper, I explore the potential benefits of applying Brunswik's (1944) concepts of EV and representative design to UXE. I present ECOVAL, a framework intended for increasing the external validity of UX experiments by allowing HCI researchers to assess and manipulate the ecological validity of cues and to achieve representative design. I developed ECOVAL primarily to help HCI researchers justify their choices made in designing experiments and avoid overgeneralizing their findings in reporting UXE findings.

ECOVAL is, I believe, a significant contribution to HCI for three reasons. Firstly, it clarifies the EV concept. Second, it provides HCI researchers with an operational framework and guidelines for achieving representative design in a field where, to the extent of our knowledge, no such work has previously been conducted. Finally, by reflecting on a case study, I show how ECOVAL benefits UXE and that it warrants further application.

The paper proceeds as follows: in Section 3, I present background information on UXE and the threats to the validity of UXEs. In Section 4, I present the ECOVAL framework and its companion guidelines. In Section 5, I present an industrial case study that illustrates the usefulness of ECOVAL for increasing and discussing the external validity of UXE designs. The case also illustrates how researchers can make trade-offs that inevitably arise while conducting UXEs between experimental validity and UX resources. In Section 6, I discuss avenues for future research and challenges that future work should address. Finally, in Section 7, I conclude the paper.

3 Background

3.1 User Experience Evaluation

UX refers to "a person's perceptions and responses that result from the use or anticipated use of a product, system or service" (ISO, 2009). Further, one can break down the high-level constructs of UX (aspects such as aesthetics, emotions and usability) into criteria or characteristics that one evaluates through metrics (Table 1).

Table 1. Examples of UX Constructs, Criteria, and Metrics

Constructs	Criteria	Metrics
Aesthetics (Zen & Vanderdonckt, 2014)	Alignment Balance Density Symmetry	Image processing or subjective expert inference
Emotions (Agarwal & Meyer, 2009; Mauss & Robinson, 2009)	Experience Physiology Verbal behavior Visual behavior Facial behavior	Self-report ("wow" experience, identification) Physiological response (heart rate, EEG, MRI, etc.) Pitch, amplitude, prosody Eye-tracking data (dwell time, fixations, hit ratio, etc.) Facial recognition or observer subjective rating
Usability (Nielsen, 2012)	Learnability Efficiency Memorability Errors Satisfaction Problems	Performance over trials (task time/completion, errors) Performance (completion rate, task time, errors) Performance over trials, self-report Errors (type, number, frequency) Self-report Subjective expert inference (e.g., heuristic evaluation)

There are three classes of UXE methods: analytic, empirical, and self-reported. Analytic methods such as heuristic evaluation or cognitive walkthroughs typically involve experts' using their knowledge of users and technology to predict usability problems. Empirical methods such as the think-aloud method or user testing usually involve representative users' executing tasks with a computer system representation to measure UX metrics directly. Self-reported methods such as survey research or experience sampling involve asking participants about their feelings, attitudes, beliefs, and so on. In this paper, I focus on empirical studies.

3.2 Threats to the Validity of UXEs

The threats to the validity of UXEs include cause-effect issues (e.g., statistical conclusion and internal validity) and generalization issues (e.g., construct validity and external validity). Statistical conclusion validity refers to the validity of conclusions about whether the observed covariation between variables results from chance, and it concerns sources of error and the appropriate use of statistical tests for dealing with such errors (Cook, Campbell, & Peracchio, 1990). Internal validity reflects whether the experimental variable or a third variable that correlates with both dependent and independent variables caused the change in the dependent variable (effect), and it concerns bias (Cook et al., 1990). Construct validity refers to the validity with which the researcher labels cause-and-effect operations in theory-relevant or generalizable terms (Cook et al., 1990). It focuses on the quality of experimental manipulations and measurements, which Gray and Salzman (1998) refer to as causal construct validity and effect construct validity, respectively. External validity reflects whether one can generalize the causal relationship beyond the experimental instance to other persons, settings, and times (Cook et al., 1990). In particular, it asks the question: "To what populations, settings, treatment variables, and measurement variables can this effect be generalized?" (Campbell & Stanley, 1966).

Table 2 depicts the threats to experimental validity discussed in HCI (Gray & Salzman, 1998).

Table 2. Threats to the Validity of UXE Studies

Statistical conclusion validity	
Statistical power	Problem: oversight of true differences Solution: increase the number of participants
Fishing and error rate problem	Problem: type I error when multiple comparisons are made Solution: test of Tukey or Scheffe, multivariate analysis of variance
Reliability of measures	Problem: test-retest reliability Solution: use longer tests or decrease interval between tests
Reliability of treatment implementation	Problem: differences in the way the treatment is implemented Solution: make the treatment as standard as possible (automation)
Random irrelevancies in the setting	Problem: extraneous source of variation in the setting Solution: controlled experiment (lab)
Heterogeneity of population	Problem: differences between individuals affect dependent variable Solution: select homogeneous population; block on characteristics more highly correlated with dependent variable; choose within-subject error terms
Internal validity	
History	Problem: extraneous event between a pretest and a posttest Solution: decrease interval between tests
Maturation	Problem: neglect of natural developmental changes (e.g. getting tired) Solution: keep duration of sessions reasonable
Testing	Problem: better results due to previously having taken a test Solution: post-test only or random assignment
Instrumentation	Problem: inconsistent measurement of dependent variable Solution: automated data collection, same instruments/observers
Statistical regression	Problem: groups defined on the basis of their extreme pretest scores Solution: increase pretest reliability
Selection	Problem: effect due to individual differences between groups Solution: random assignment
Mortality	Problem: unequal dropout rates among comparison groups Solution: one-group design or random assignment
Construct validity	
Poor construct definition	Problem: inaccuracies and errors in the construct definition Solution: comply with construct terminology
Mono-operation bias	Problem: single treatment used Solution: use multi-group design or pretest
Mono-method bias	Problem: single measure used Solution: use several methods (e.g. user tests combined to questionnaires)
Interactions	Problem: between treatments or between testing and treatment Solution: good planning and monitoring of the subjects
Construct confounding	Problem: confounding levels of constructs with constructs Solution: conduct parametric research involving several levels of constructs
Social threats	Problem: hypothesis guessing, evaluation apprehension, experimenter bias Solution: keep social interactions to a minimum
External validity	
People	Problem: unrepresentativeness of sample population Solution: random selection of participants
Setting	Problem: unrepresentativeness of experimental setting Solution: replicate experiment in different settings
Time	Problem: peculiar time the experiment took place Solution: replicate experiment at different times

Table 3 summarizes the sources of invalidity for a selection of UXE methods and covers the major experimental designs and evaluation methods employed in HCI. In particular, formative and summative usability testing are evaluation methods broadly adopted in HCI. Formative usability is an iterative test-and-refine method applied early in the design process and usually involves a within-subject design with few participants. In contrast, summative usability is a singulative quality-insurance method applied later in the design process and usually involves a between-subjects design with a larger sample of participants. Formative usability supports decision making during product, whereas summative usability is a tool for describing the UX (Tullis & Albert, 2013).

Table 3. Sources of Invalidity of UXE Methods

UXE methods	Analytic methods (one-shot analysis)	Single-group think-aloud	Single-group case study	Within-subject formative usability	Within-subject lab experiment	Between-subjects summative usability	Between-subjects A-B testing	Counterbalanced lab experiment	Longitudinal survey research	Longitudinal experience sampling
Statistical power	–	–	–	?	+	?	+	?	?	?
Heterogeneity of population	–	?	?	+	+	?	+	?	+	+
Maturation	–	–	–	–	–	+	+	+	–	–
Testing			?	–	–	+	+	+	–	–
Instrumentation	–		?	+	+	+	+	+	?	?
Selection	–	–	+	?	?	+	+	+	+	+
Mortality		–	+	–	–	+	+	+	–	–
Construct confounding		?	?	?	?	?	?	?	?	?
People		–	?	+	+	?	?	?	?	?
Setting			?	?	?	?	?	?		?
Time							?			?

Note: a minus indicates a definite weakness, a plus indicates that the factor is controlled, a question mark indicates a possible source of concern, and a blank indicates that the factor is not relevant.
Within subject: single-group design, repeated measures; between subjects: multiple-group design, one measure.
True A-B tests require between-subjects design and random assignment between groups A and B.
Longitudinal designs involve repeated measures of the same random sample.

As Table 3 shows, UXE methods that increase internal validity tend to jeopardize external validity and vice versa. Furthermore, UXE methods strongly emphasizes the generalization to organism (people) at the expense of both experimental setting and time. Finally, UXE methods never fully logically justify external validity (see question marks). To quote Campbell and Stanley (1966, p. 5): "Internal validity is the basic minimum without which any experiment is uninterpretable…. While internal validity is the sine qua non, and while the question of external validity…is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal.".

4 The ECOVAL Framework

The ECOVAL framework (Figure 2) is an adaptation of Brunswik's lens model (Figure 1) to the HCI field. It keeps Brunswik's concepts of ecological validity and representative design intact while capturing the concepts of HCI that are the most relevant to UX evaluations.

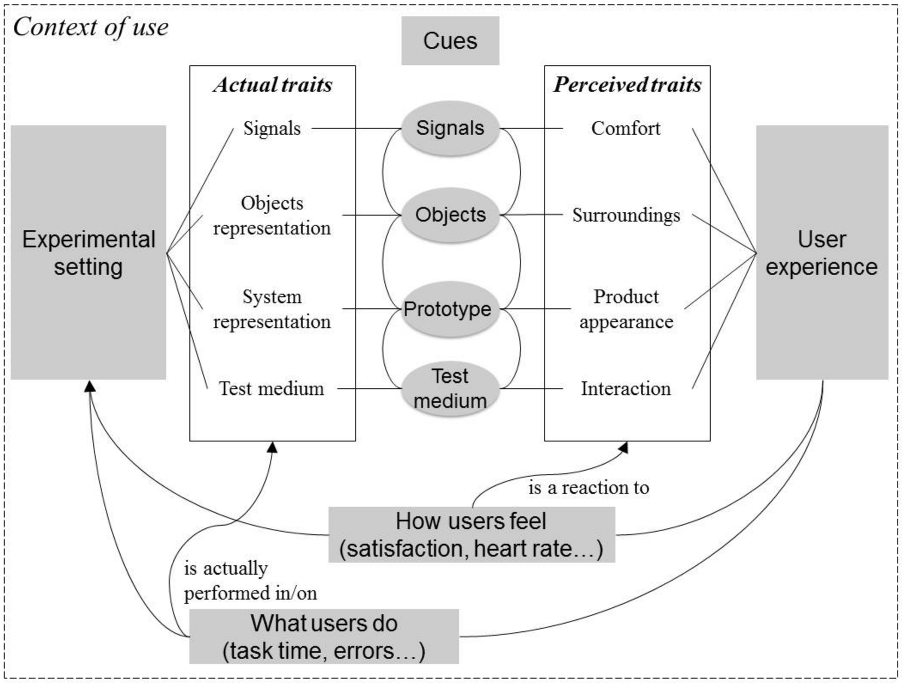


Figure 2. The ECOVAL Framework

The ECOVAL framework breaks down the actual state of the environment into actual traits that correspond to perceived traits. Actual traits are the objective characteristics of the cues present in the environment. Perceived traits correspond to the subjective characteristics of these cues, which users exploit to infer the product's state, to develop a mental model for how it should work, and to apply a response to use it, which complies with earlier work about the action cycle (Norman, 1983, 1984). Users' motor and emotional responses (i.e., what they do and how they feel) reflect the UX in the experimental setting. UX metrics such as performances or satisfaction measure users' achievements, which complies with earlier work about user performance and perceived usability (Hassenzahl, 2004; Sonderegger & Sauer, 2010) and the context of use (Shackel, 1991). Four classes of cues are actually present in the environment and contribute to users' attitude toward a product and the experimental setting: signals, objects (other than the product being tested), prototype, and test medium (i.e., the physical device).

The signals refer to stimuli such as noise, temperatures, or lighting that one can objectively measure. For example, one can express noise in decibels and temperature in Celsius degrees (see Figure 3). One can also quantify human perceptions of noise and temperature. Individuals can experience noise in a spectrum that ranges from barely audible to painfully loud. Similarly, individuals can experience temperatures as comfortably low or harmfully high.

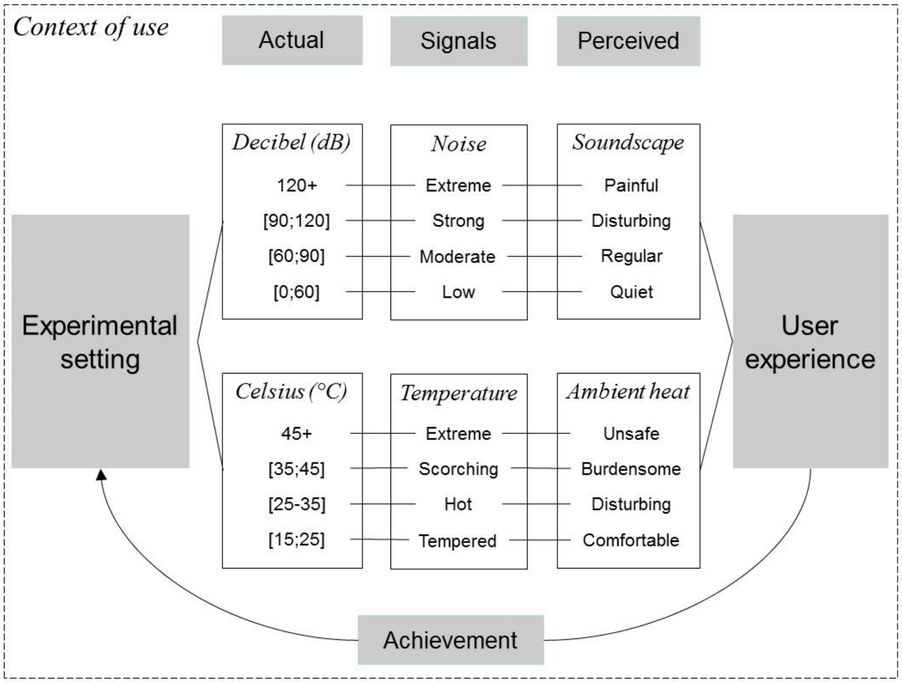


Figure 3. Examples of Cues in the Signals Class

The prototype is the computer system representation, which one can characterize in terms of visual refinement, dynamicity, data model, breadth, and width of functionality (McCurdy, Connors, Pyrzak, & Kanefsky, 2006). Visual refinement refers to the representative state of the product (i.e., final product, wireframe, or sketch). Dynamicity refers to whether the product has been implemented, simulated, or neither. One can also quantify human perceptions of visual refinement and dynamicity: from concrete to abstract for visual refinement or from dynamic to static for dynamicity (see Figure 4).

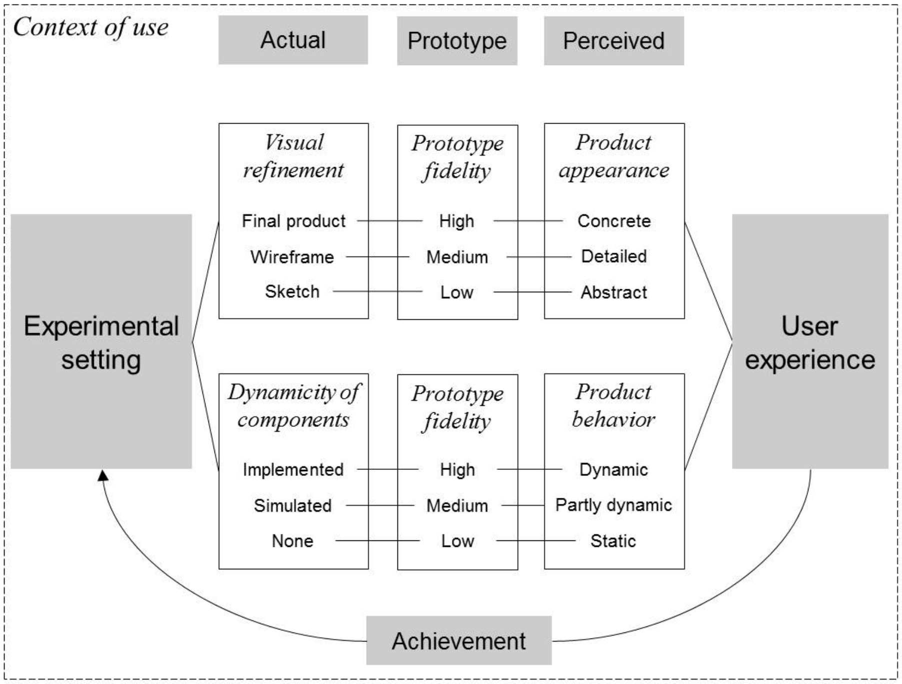


Figure 4. Examples of Cues in the Prototype Class

Table 4 shows cues and, for each cue, the correspondence between actual and perceived traits.

Table 4. Cues, Actual Traits, and Perceived Traits

Cues	Actual traits	Perceived traits
Signals (noise)	Decibels (dB) • 120+ • [90;120] • [60;90] • [0;60]	Soundscape • Extreme, painful • Strong, disturbing • Moderate, regular • Low, quiet
Signals (temperature)	Celsius (°C) • 45+ • [35;45] • [25;35] • [15;25]	Ambient heat • Extreme, unsafe • Scorching, burdensome • Hot, disturbing • Tempered, comfortable
Objects (type)	Representation • Original • Copy (replica) • Simplified (mock object)	Surroundings • Everyday objects • Similar objects • Simplified objects
Objects (building)	Testing location • Field (uncontrolled) • Laboratory (controlled)	Testing environment • Natural (workspace) • Artificial
Prototype (refinement)	Visual refinement • Real system • Wireframe or sketch • Presentation or video	Product appearance • Concrete, real • Detailed, simplified • Abstract, conceptual
Prototype (dynamicity)	Dynamicity of components • Implemented • Simulated • Not implemented	Product behavior • Dynamic, smooth • Partly dynamic, fragmented • Static, rigid
Prototype (data model)	Data model (database) • Real • Sample • Randomly generated	Product dataset • Complete • Uncomplete • Irrelevant or unrepresentative
Prototype (breadth)	Completion (%) • [80;100] • [50;80] • [0;50]	Product completion • Final product • Prototype • Concept
Prototype (width)	Completion (%) • [80;100] • [50;80] • [0;50]	Product completion • Final product • Prototype • Concept
Test medium (device)	Tool • Computer • Paper-and-pencil • None	Product palpability • Tangible with continuous interaction • Tangible with fragmented interaction • Intangible with envisioned interaction
Test medium (interaction)	Input modalities • Direct manipulation • Indirect manipulation • Oral command	Interaction with the product • Performed for real (log-files) • Mimicked with a pencil (annotation) • Verbalized (comments)

4.1 Ecological Validity of Cues in User Experience Evaluations

The EV of cues in UXEs refers to the correlation between perceived traits and UX metrics. The cues have an EV close to 1 when perceived traits are highly correlated with UX metrics and an EV close to -1 when perceived traits are highly negatively correlated with UX metrics. They have an EV close to 0 when perceived traits and UX metrics are not correlated.

The signals present in the environment (e.g., level of noise) highly correlate with the level of sensory comfort that users perceive. As a consequence, signals usually have an EV close to 1. Similarly, objects usually have an EV close to 1 because artefacts present in the testing environment highly correlate with users' ability to achieve their goals. Similar reasoning holds for the prototype and the test medium because the product’s appearance and palpability are highly correlated with the user experience. Considered independently, the cues seem to have an EV close to 1 (Table 4).

4.2 Inter-correlations of Cues

Unfortunately, estimating the EV of a product as a whole is not as straightforward as estimating the EV of each cue independently. The interrelatedness of the product's proximal cues (Table 4) introduces redundancy or inter-correlations into the environment (Brunswik, 1952). For example, the prototype is an abstract representation of software (bits, data structures, algorithms, etc.), whereas the test medium is a physical device that allows interaction. The inter-correlations between visual refinement, dynamicity, tool, and input modalities give the product a specific “look and feel” in terms of appearance, behavior, palpability, and interactivity. These inter-correlations significantly affect the task representativeness and, therefore, the outcome of UXEs.

I believe that ECOVAL allows researchers to predict how the inter-correlations of cues affect the UX. In particular, ECOVAL helps one to predict how specific combinations of cues affect an organism’s responses and behaviors (Dhami et al., 2004) and how specific combinations of cues enable users “to perceive in order to act, but also to act in order to perceive” (Araujo et al., 2007). Nevertheless, we need to gain a better understanding of the cues that users perceive and use. Considering the extremely large number of all possible combinations of cues, the case study reported here primarily focuses on the interrelation between visual refinement, dynamicity, tool, and input modalities.

4.3 The ECOVAL Guidelines

I produced the proposed guidelines for using ECOVAL as a seven-step procedure:

1. Conduct task analysis if one needs domain-expertise and/or extended knowledge about user task
2. Set study goals and choose UX metrics
3. Identify cues that the experiment will represent
4. Define experimental design
5. Specify ideal instance of experimental plan (e.g., experimental conditions, sampling method, random assignment, material, etc.), assess its feasibility against organizational constraints such as safety or availability of participants, and adjust if necessary
6. Assess experimental validity of experimental plan (Table 3 may serve as baseline assessment tool; one may modify the + and – indicators according to experimental plan specificities) and proceed to trade-offs if necessary (e.g., smaller sample size with stratified sampling or more task instances with fewer participants), and
7. Conduct experiment, analyze results, and report findings.

5 Case Study

The case study took place in a company whose core-business is hot-dip galvanizing (GA), which involves applying a zinc coating on steel products by immersing them in a bath of molten zinc in order to protect the underlying steel from corrosion. The sheet is continuously fed through a cleaner, an annealing furnace, and a molten zinc bath (Dallin, 2005).

The UX team had to develop a prototype for a mobile system for the monitoring activities in order to increase both the productivity and organizational efficiencies (better equipment traceability, reduction of unplanned downtime, and increased production quality). The UX team followed the usability engineering (UE) lifecycle (Mayhew, 1999), which involved contextual task analysis, work reengineering, informal evaluation of user stories, screen design standards, and formative usability testing. Formative usability involved three iterations: one paper-and-pencil evaluation and two rounds of user testing (Table 5). Raphaël Schramme (MS student) and I composed the UX team throughout the case study. Ugo Braga

Sangiorgi (PhD student) joined us from steps 5 to 7. Mathieu Zen (PhD student) joined us for the step 7. In this section, I report how the UX team followed the ECOVAL guidelines in the case study.

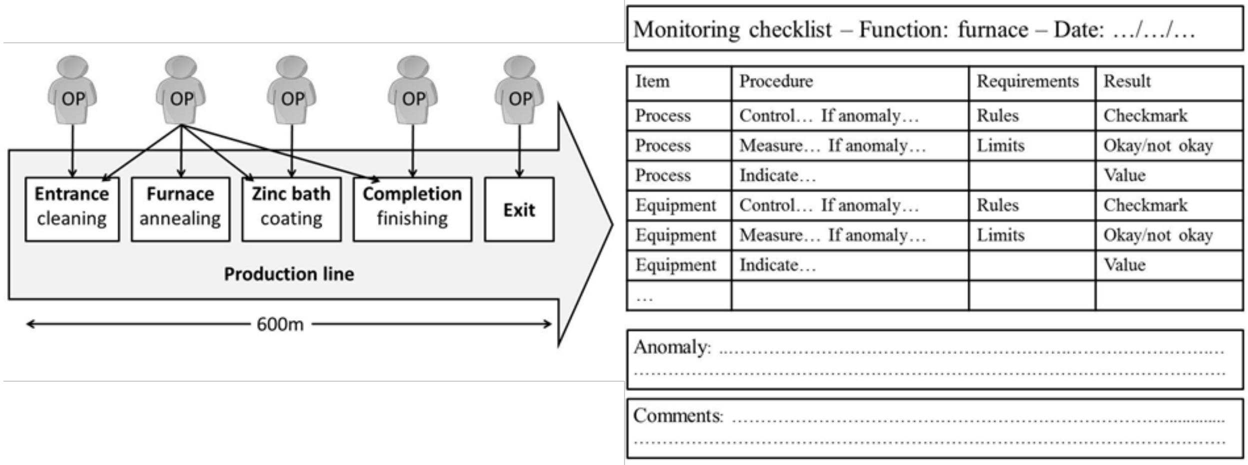
Table 5. UX Plan

Method	Scope and outcome
Requirements analysis	
Contextual task analysis	Task characteristics, user problems, usability goals Work models: sequence, physical, and cultural
Work reengineering	Task model and associated user stories
Informal evaluation	Interview with 3 user representatives (managers) Validation of the user stories
Design and prototyping	
Screen design standards	Production of screens with a wireframing tool
Formative usability testing	
Iteration 1	Paper-and-pencil evaluation Wireframing screens printed on paper sheets
Iteration 2	User testing Wireframing screens running as interactive prototypes
Iteration 3	Comparative experiment Prototypes: paper mock-up versus interactive prototype

5.1 Guideline 1: Conduct Task Analysis

Using contextual task analysis (CTA), the UX team could gain insight into the work organization and the user tasks involved in monitoring activities. CTA involved gathering data about users' physical environment (signals and objects), identifying the main work artifacts and objects, collecting task scenarios, and gaining insights into users' problems, bottlenecks, and errors.

I summarize the findings as follows. With around 300 items checked per day (i.e., 90,000 items each year), the monitoring activity is critical for the organization. Four teams of operators (i.e., 32 people in total) work in rotating shifts of eight hours: morning, afternoon, and night shifts, plus one team at rest. The production requires a minimum of five operators (Figure 5, left). Every day operators are requested to fill in a pen-and-paper monitoring checklist. A specific checklist is assigned to each operator according to the operator's posting on the GA line. Each checklist is structured as a four-column table (Figure 5, right): from left to right, item, procedure, requirements and result. The checklist also includes an anomaly and comments section. Both anomalies and comments have to be described in a computer system that operators unanimously perceive as "a waste of time", "a nuisance", "complicated to use", and "irrelevant for the job".



Monitoring checklist – Function: furnace – Date: …/…/…

Item	Procedure	Requirements	Result
Process	Control… If anomaly…	Rules	Checkmark
Process	Measure… If anomaly…	Limits	Okay/not okay
Process	Indicate…		Value
Equipment	Control… If anomaly…	Rules	Checkmark
Equipment	Measure… If anomaly…	Limits	Okay/not okay
Equipment	Indicate…		Value
…			

Anomaly: ………………………………………………………………………………
………………………………………………………………………………………

Comments: ……………………………………………………………………………
………………………………………………………………………………………

Figure 5. Work Organization (Left) and User Task (Right) (OP: Operator)

5.2 Guideline 2: Set Study Goals and Choose UX Metrics

With the information provided through CTA, the UX team could produce work models, task models, user stories, and screen design standards (Table 5). The UX team created the screen design standards with Balsamiq (2008). The UX team then implemented Balsamiq screens either as mock-ups printed on paper or as interactive prototypes by interlinking them in GAMBIT (Figure 6). The UX team chose GAMBIT (Sangiorgi & Vanderdonckt, 2012) as the supporting tool for user testing because it enables one to produce interactive prototypes by simply interlinking screens, to wireframe prototypes on multiple devices, to log data unobtrusively (date and current screen displayed on device), and to make live recordings of the sequence of screens the user interacted with.

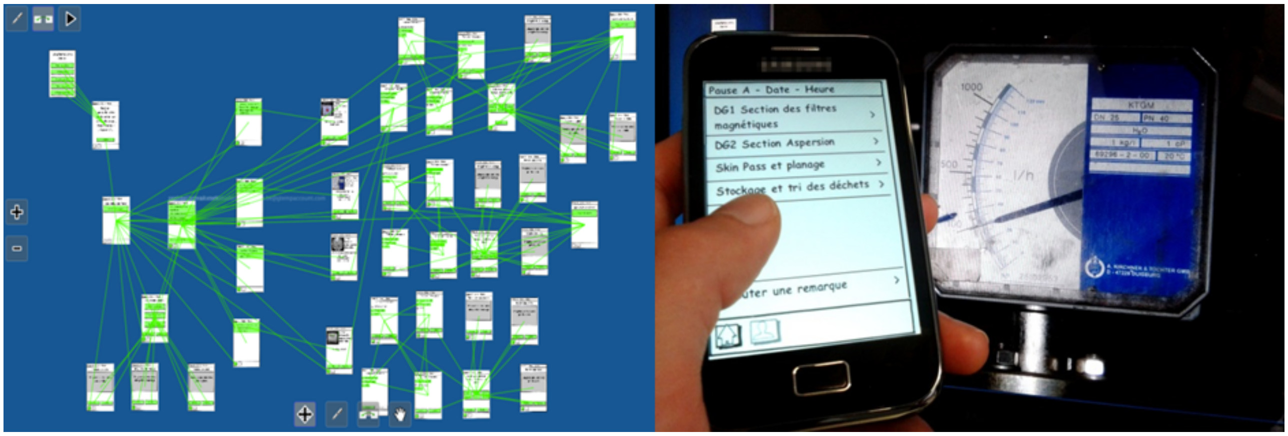


Figure 6. GAMBIT: Balsamiq Screens Interlinked to Form an Interaction Map (Left); Balsamiq Screens Running as Interactive Prototype on a Mobile Device (Right)

As Table 5 shows, the UX team conducted three rounds of UXE during the formative usability testing phase. Table 6 shows our study goals and UX metrics.

Table 6. Study Goals and UX Metrics

Iteration	Study goals	UX Metrics
1	Detection of usability issues	User errors
2	Detection of usability issues Analysis of user efficiencies Analysis of user satisfaction	User errors Task time and number of screen looked through Self-reported user satisfaction
3	Analysis of user preferences	Self-reported preferences

As Tullis and Albert (2013) advocate, the UX team checked the following errors to detect usability issues:

1. Behaviors that prevent task completion (type 1)
2. Mistaken belief that a task is completed when it actually is not and vice versa (type 2)
3. Oversight of something that should be noticed (type 3), and
4. Misinterpretation of some piece of content (type 4).

In order to detect usability issues, the UX team first performed the paper-and-pencil evaluation (iteration 1). The results from this evaluation indicated no usability problems. In fact, all participants spontaneously expressed very positive judgments about the design. They stated that the navigation that they envisioned with the product seemed easy, fast, and comfortable. Therefore, I implemented the Balsamiq material in a GAMBIT prototype “as is”.

Then, in order to study users’ efficiency and satisfaction while navigating the user interface (UI), the UX team performed the user testing evaluation (iteration 2). The UX team measured the users’ efficiency based on the duration and number of navigated screens to complete each task. The UX team extracted this information from the log file in terms of the time elapsed and the number of screens looked through to get a task done. The UX team used a computer system usability questionnaire (CSUQ) with a five-point rating scale to collect self-reported rates of usability (Lewis, 1995). The CSUQ rates the usability of computer systems in terms of system usefulness (SYSUSE), information quality (INFOQUAL), interaction quality (INTERQUAL) and overall.

Finally, the UX team carried out the comparative experiment (iteration 3) in order to compare the users' preferences between a paper and a GAMBIT prototype.

5.3 Guideline 3: Identify Cues to be Represented in Experiment

With the information provided through CTA, the UX team designed an experimental task that captured the cues relevant for achieving task representativeness and, therefore, representative design. The key aspects of users' task involved items, procedures, and requirements. The cues relevant for achieving task representativeness included the representation of the actual state of each item and the tool supporting the execution of the monitoring activities. The experimental task involved the following steps:

1. Select item from list
2. Locate and reach item in the environment
3. Read monitoring procedure and requirements associated with item
4. Check whether requirements were satisfied, and
5. Enter result; if requirements not satisfied, report anomaly.

Further, with the information provided through CTA, the UX team identified the cues that captured the key aspects of the environment. The cues that captured the key aspects of the environment included the level of noise, the ambient temperature, and the items to check.

The UX team extracted and selected all items involved in the UXE from the formal paper checklist so that any user would recognize and be familiar with them. By enabling users "to perceive in order to act, but also to act in order to perceive", the task and the cues represented in the experiment allowed action fidelity (Stoffregen et al., 2003); that is, behaviors in an experimental context that reproduce those in the intended environment.

5.4 Guideline 4: Define Experimental Design

The UX team chose a within-subject design for iterations 1 and 2 and a between-subjects with Latin Square design for iteration 3. The UX team applied stratified sampling across postings to select a representative sample of the population, which allowed the UX team to increase external validity while maintaining defensible internal validity. Participants involved in iteration 2 were first-time users (i.e., users who were not involved in iteration 1). Table 7 describes the population sample.

Table 7. Population Sample

Iteration	Size	Profiles
1	10	10 operators (2 per posting)
2	18	15 operators (3 per posting) + 3 managers
3	10	10 operators (2 per posting)

5.5 Guideline 5: Specify, Assess and Adjust Ideal Instance of Experimental Plan

Initially, the ideal experiment the UX team envisioned had the following characteristics:

- Number of experimental tasks: each participant would perform the experimental task (guideline 3) 10 times, which would increase the statistical conclusion validity, and
- Testing location: participants would perform the tasks in the field and in the vicinity of the items so the UX team could capture the relevant aspects of the real world and, therefore, increase the external validity.

Unfortunately, these points were difficult and dangerous to realize. Firstly, the work environment was uninviting due to the lack of natural light, the amount of strong noise and dust, and the high risk of accidents. To protect themselves from these unfriendly conditions, both the workers and the researchers needed to wear earplugs, safety shoes, helmets, and coveralls at all times at all places along the production line. As such, the UX team could not safely execute the ideal scenario for our experiment.

Second, that ideal scenario would have put an excessive demand on the users' time and effort given the fact that the work organization was complex and involved multiple concurrent functions and distributed workers in space (Figure 5). Third, the scenario would have resulted in a number of experimental conditions equal to the number of postings (5) and, thereby, threatened both internal and external validity.

Given these issues, the UX team made trade-offs between the ideal experiment, organizational constraints, and UX resources but not at the expense of experimental validity. First, the UX team set the number of experimental tasks at four per participant (T1 to T4), which helped to maintain a reasonable duration of the session (30 minutes in total including instructions, test, and debriefing) and was consistent with the four types of task identified during CTA (i.e. control, locate, indicate and measure)

The second trade-off involved adjusting the testing location to alternate between evaluations that were less and more demanding in terms of UX resources. The paper-and-pencil evaluation (iteration 1) took place in the field because it involved 10 participants, only one researcher, and no GAMBIT development. The GAMBIT evaluation (iteration 2) took place in the lab because it involved 18 participants, three researchers, and GAMBIT development that required UX resources. The comparative experiment (iteration 3) took place in the field again because it involved 10 participants, only one researcher, and light GAMBIT development.

The UX team introduced mock objects in iteration 2 in an attempt to "bring the environment into the lab" and, therefore, increase action fidelity. These mock objects were the pictures of the objects mentioned by the items on the checklist, each displayed on a separate 21-inch monitor. Mock objects helped us increase the action fidelity by forcing the participants to walk from one monitor to another in order to perform the four tasks.

In the end, this experimental setup (Figure 7) represented a safer setting for both the participants and the researchers. Furthermore, it allowed the UX team to strike a good balance between internal validity (i.e., the experimental condition was the same for each participant) and action fidelity (i.e., the experimental behavior reproduced the behavior in the intended environment).

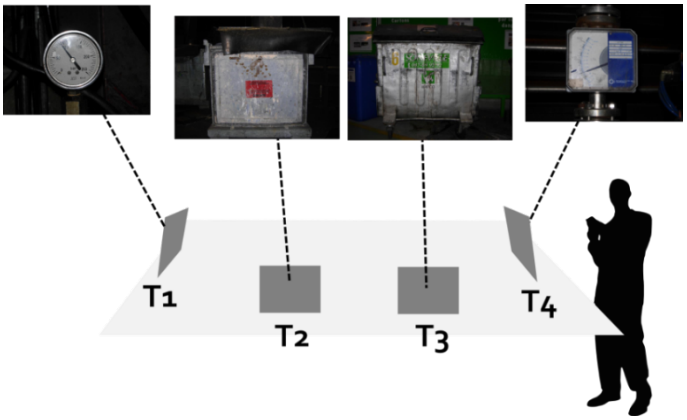


Figure 7. Mock Objects Representing Items to Check

Surprisingly, the action fidelity was even higher in these lab conditions than in the field because all participants went to actually check the equipment. In the field, they remained seated at the table and imagined their responses (Tables 8 and 9).

In iteration 3, the UX team studied the influence of dynamicity, device, and interaction on task representativeness. The UX team used the same cues in the signals and objects classes: noise, temperature, testing location, and items. These cues were representative of the users' natural setting. Furthermore, iteration 3 also involved the same visual refinement, data model, and completion rate. The paper condition and the GAMBIT condition differed only for three cues: the dynamicity of the prototype (respectively not implemented versus implemented), the device (respectively paper-and-pencil versus smartphone), and the interaction (respectively mimicked versus performed for real). Neither of these two experimental settings achieved action fidelity because the participants remained seated at the table to invent their response.

The behavior of the participants during iteration 3 strongly suggests that field experiments do not guaranty task representativeness. On the contrary, lab experiments that capture relevant cues via mock objects may be more likely to increase task representativeness.

Table 8. Values and Representativeness of Cues in Iterations 1 and 2

Cues	Values		Representativeness	
	Iteration 1	Iteration 2	Iteration 1	Iteration 2
Noise	[90;120] dB	[0;60] dB	+	–
Temperature	[25;35] °C	[15;25] °C	+	–
Testing location	Production line	Laboratory	+	–
Items (objects)	Real objects	Mock objects	+	+
Visual refinement	Wireframe	Wireframe	+	+
Dynamicity	Not implemented	Implemented	–	+
Data model	Sample (2%)	Sample (4%)	–	–
Completion	[0;50] %	[0;50] %	–	–
Device	Paper-and-pencil	Smartphone	–	+
Interaction	Mimicked	Performed for real	–	+
Action fidelity	No	Yes	–	+
Note: a minus indicates a low representativeness; a plus indicates a high representativeness. Iteration 1: paper-and-pencil evaluation; iteration 2: user testing evaluation with GAMBIT.				

Table 9. Values and Representativeness of Cues in Iteration 3

Cues	Values		Representativeness	
	Paper	GAMBIT	Paper	GAMBIT
Noise	[90;120] dB		+	
Temperature	[25;35] °C		+	
Testing location	Production line		+	
Items (objects)	Real objects		+	
Visual refinement	Wireframe		+	
Dynamicity	Not implemented	Implemented	–	+
Data model	Sample (8%)		–	
Completion	[0;50] %		–	
Device	Paper-and-pencil	Smartphone	–	+
Interaction	Mimicked	Performed for real	–	+
Action fidelity	No		–	
Note: a minus indicates a low representativeness; a plus indicates a high representativeness.				

5.6 Guideline 6: Assess Internal and External Validity of Experimental Plan

Table 10 represents the experimental validity of the experimental plan.

Table 10. Experimental Validity

Threats to Experimental Validity	Iteration 1	Iteration 2	Iteration 3
Statistical power			+
Heterogeneity of population	+	+	+
Maturation	+	+	+
Testing	+	+	–
Instrumentation	–	+	–
Selection	+	+	+
Mortality	+	+	+
Construct confounding	+	+	+
People	+	+	+
Setting	**+**	**–**	**+**
Task (Action fidelity)	**–**	+	–
Note: a minus indicates a definite weakness, a plus indicates that the factor is controlled, a question mark indicates a possible source of concern, and a blank indicates that the factor is not relevant.			

5.6.1 Statistical Conclusion Validity

The UX team controlled the statistical conclusion validity well. Statistical power is not a concern of formative usability, which usually involves between five and eight participants (Nielsen & Landauer, 1993). The sample size exceeded recommendations for this UXE method (Table 7). The UX team made such an expensive choice in order to maintain good social relationships with the workforce and avoid inter-team competition. Because of the stratified sampling, the heterogeneity of the population was not a concern.

5.6.2 Internal Validity

The UX team controlled the internal validity best in iteration 2. Specifically, the duration of the test did not exceed 20 minutes per participant (maturation), none of the participants had previously taken the test (testing), the UX team measured the dependent variables consistently (i.e., automated data collection (instrumentation)), the UX team used stratified sampling across postings (selection), and none of the participants dropped out (mortality).

5.6.3 Construct Validity

The UX team controlled the construct validity appropriately. Indeed, the UXE methods were relevant to the goals of each study, the UX team properly defined the combination of cues (Tables 8 and 9), and the UX team made the measurements in compliance with prior research (Mayhew, 1999; ISO, 2009; Tullis & Albert, 2013). Moreover, the UX constructs and metrics employed in the experiments are widely acknowledged and well defined, which helped the UX team to avoid problems of interpretation of the findings (effect construct).

5.6.4 External Validity

The UX team controlled the representativeness of the sample population via stratified sampling. The representativeness of the setting was higher in iteration 1 and 3 than in iteration 2. However, the behaviors observed in the lab context of iteration 2 reproduced those in the intended environment, while those observed in the field context of iterations 1 and 3 did not. The mock objects introduced in iteration 2 increased action fidelity (Table 8) so the UX team could generalize the findings of iteration 2 to monitoring activities.

5.6.5 Generalization of the Findings

These results hold for technicians, team leaders, and, to a lesser extent, production managers because these other user groups have the same characteristics as the operators and perform a similar task in the same environment. The generalization of the findings to other populations in the same organization further highlights the usefulness of CTA for achieving representative design. In fact, the UX team could only make these claims thanks to the extensive understanding of the organizational culture that the UX team gained through the CTA.

5.7 Guideline 7: Conduct Experiment, Analyze Results, and Report Findings

5.7.1 Iteration 1: Paper-and-pencil Evaluation

The UX team detected no usability problems. On the contrary, all participants spontaneously expressed very positive judgments about the design. Accordingly, the UX team implemented the Balsamiq material "as is" in a GAMBIT prototype to conduct iteration 2.

5.7.2 Iteration 2: GAMBIT Evaluation

The first six participants consistently made the same error: they believed that the task was not completed when it actually was (type 2). This error resulted from their oversight of a validation button (type 3) or a misinterpretation when reaching the homepage (type 4). During the debriefing, they reported to have experienced difficulties in navigating the UI because they found it absurd to have to confirm their inputs. Hence, before resuming the evaluation with the remaining 12 participants, the UX team modified the prototype in accordance with these suggestions, which meant altering 20 percent of the screen design and re-interlinking it in a second prototype.

Participants performing with the second prototype made no such error. During the debriefing, the participants focused almost exclusively on their UX with the future system. Their feedback related to workday organization (when to launch the monitoring), operational organization (where to put the device in charge), or extended uses of the system (how it could support other activities). The participants' efficiency increased with the second prototype: it only took 20 seconds and 3.2 screens on average to complete a task compared to 35 seconds and 8.6 screens on average with prototype 1. The UX team also saw an increased satisfaction among the participants when using prototype 2 (Table 11).

Table 11. Score Comparison Between Prototypes 1 and 2

	Prototype 1			Prototype 2		
	Mean 1	Low 1	High 1	Mean 2	Low 2	High 2
SYSUSE	4.29	3.73	4.85	4.49	3.96	5.00
INFOQUAL	3.81	3.25	4.36	4.40	3.87	4.93
INTERQUAL	3.89	3.17	4.60	4.28	3.68	4.88
OVERALL	4.33	3.89	4.78	4.33	3.89	4.78

High SYSUSE and OVERALL scores for both prototypes indicate a high level of satisfaction regarding the system usefulness and the UX with the system. Both INFOQUAL and INTERQUAL scores were much higher for prototype 2, which indicates that the information was better organized in this prototype and that the participants preferred to interact with it as compared to prototype 1.

5.7.3 Iteration 3: Comparative Experiment

All participants strongly expressed their preference for the GAMBIT evaluation. They reported that using a computer system made it easier for them to "project themselves into real use" and to feel "in charge" and "proactive". They also stated that it was "less intrusive" compared to paper-and-pencil evaluations during which the researcher "flips the sheets of paper as the interaction runs through".

5.8 Summary of the Results

Formative usability involving an interactive prototype (iteration 2) enabled our detecting usability problems that remained overlooked during paper-and-pencil evaluation (iteration 1) even though both UXEs used the same visual refinement. An empirical experiment conducted in the field in which the UX team compared two test medium instances (iteration 3) found that the realism of the interaction was essential for both the UX and the commitment of participants to the experiment.

However, iteration 3 also highlighted that lab experiments that involve relevant cues may be more likely to increase task representativeness than field experiments. In particular, sampling the objects that are present in the real-life environment and that are relevant for users to achieve their goals seems to be more important for task representativeness than the representativeness of the prototype. This finding is consistent with earlier findings about representative design (Whitefield, Wilson, & Dowell, 1991; Sefelin, Tscheligi, & Giller, 2003).

6 Discussion

6.1 Contribution

Our findings indicate that ECOVAL presents a significant contribution to the HCI field for three reasons. First, it clarifies Brunswik's (1944) concept of ecological validity, which HCI research has inconsistently and incorrectly used. We believe that this clarification will facilitate the development of cumulative science about the EV of cues in UXEs. Second, ECOVAL provides HCI researchers with an operational framework and guidelines for applying representative design in a field where, to the extent of our knowledge, no such work has previously been conducted. In particular, the ECOVAL guidelines will help HCI researchers to design better experiments, to better describe them, and to better discuss and justify the generalization of their findings. Finally, by reflecting on a case study, I propose ECOVAL as beneficial to HCI and worthy of further application. On the one hand, the case study demonstrates that UXE tasks such as design and description of experiments are not straightforward. On the other hand, it demonstrates how essential they are to avoid the overgeneralization of experimental findings.

6.2 How to Apply the ECOVAL Guidelines

In order to address ecological validity and representative design in UXEs, future research should follow the ECOVAL guidelines. Table 12 includes what researchers should focus on during their implementation (column 2) for each of the seven guidelines (column 1) and references about related methods and techniques to properly implement them (column 3).

Table 12. How to Apply ECOVAL Guidelines

Guidelines	Focus	References
1. Conduct task analysis	Contextual enquiry	Mayhew (1999), Holtzblatt, Wendell, & Wood (2005)
2. Set study goals and UX metrics	Usability and UX	(Mayhew, 1999) (Tullis & Albert, 2013)
3. Identify cues relevant to task representativeness	Work models	Mayhew (1999), Holtzblatt et al. (2005)
4. Define experimental design	Research methods	Campbell & Stanley (1966), Trochim, Donnelly, & Arora (2015)
5. Specify ideal experimental plan, assess its feasibility and adjust	Task scenarios	Nielsen (2014)
6. Assess experimental validity and make trade-offs	Experimental validity	Table 3 Campbell & Stanley (1966)
7. Conduct experiment, analyze results and report findings	Usability and UX	Mayhew (1999), Tullis & Albert (2013)

6.3 Additional Challenges Related to EV and Representative Design in UXEs

The ECOVAL framework is a first step toward developing a clearer understanding of EV in HCI and achieving representative design in UXE studies. Moving forward, the HCI community will face several exciting challenges.

- Usability metrics (i.e., effectiveness, efficiency, and satisfaction) are not necessarily correlated (Frøkjær, Hertzum, & Hornbæk, 2000). Therefore, a given cue's EV may differ from one UX metric to another. For example, the visual refinement of the prototype may be highly correlated with user satisfaction and not correlated with users' effectiveness. Future work on UXE needs to study and report on the EV of cues according to the widest possible spectrum of UX metrics, which includes issue-based, performance, and behavioral and physiological metrics. Only then will we be able to provide recommendations about which UX metrics to use in a particular situation.
- The complete coverage of the whole population of cues present in the users' environment seems to be infeasible (Dhami et al., 2004). Therefore, achieving representative design rests in HCI researchers' ability to identify the cues and their combinations that are relevant to task representativeness (e.g., by conducting CTA). CTA significantly increases the level of effort put on the UX team while designing UXEs, especially when the tasks are collaborative and distributed in time and space. However, CTA can be supported by ambulatory assessment (AmA), which uses field methods to assess the ongoing behavior, knowledge, experience, and environmental aspects of people when executing tasks in their natural setting (Kieffer, Batalas, & Markopoulos, 2014).
- We need to further investigate the inter-correlation of cues, in particular between prototype and test medium, to provide clear-cut guidance on how to achieve representative design throughout product development. On the one hand, investigating this issue requires considering the other classes of UXE (namely, analytic and self-reported methods). On the other hand, it requires systematically sampling the EV of cues in the prototype and test medium dimensions so as to cover low- and high-fidelity prototypes.
- The consistent sophistication of information and communication technology has extended the context of use to a multi-user, multi-task, multi-platform, multi-environment paradigm. Indeed, we need to know how well the ECOVAL framework will scale across that many contexts of use. In order to tackle this issue, we need further cases studies that involve collaborative tasks or multi-platform settings.

7 Conclusion

In this paper, I identify inconsistencies and inaccuracies regarding the application of Brunswik's (1944) concepts of ecological validity and representative design in the current HCI literature. I argue that HCI researchers can use the ECOVAL framework to design valid UX experiments provided that they have collected sufficient information about users' tasks and environment prior to the design phase. I also list guidelines for measuring the EV of cues and achieving representative design. The primary focus of these guidelines is to increase task representativeness by bringing the environment into the lab through mock objects and by manipulating the inter-correlations between prototype and test medium. I report on a case study as a proof of concept to demonstrate the relevance of ecological validity and task representativeness to UXE (and, hence, UX design). Finally, I also propose this step-by-step procedure to help HCI researchers design valid UX experiments, discuss experimental validity in a structured manner, and understand why claims made about generalization of UX findings can have limited accuracy. Future research should provide more detailed guidance for inter-correlations and trade-offs between the ideal experimental setup and actual organizational constraints in UXE. Further case studies will need to affirm the generalizability of the approach for addressing different work contexts and different needs of the UX team.

Acknowledgments

I acknowledge the support of this research by the project Virtuoso (Plan Marshall) under Convention n°6608 and by the project QualIHM under Convention n°1217570, both funded by DGO6, Région Wallonne, Belgium.

References

Agarwal, A., & Meyer, A. (2009). Beyond usability: Evaluating emotional response as an integral part of the user experience. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.*

Araujo, D, Davids, K. W., & Passos, P. (2007). Ecological validity, representative design, and correspondence between experimental task constraints and behavioral setting: Comment on Rogers, Kadar, and Costall (2005). *Ecological Psychology*, *19*(1), 69-78.

Balsamiq. (2008). Retrieved from https://www.balsamiq.com/

Bernstein, M. S., Ackerman, M. S., Chi, E. H., & Miller, R. C. (2011). The trouble with social computing systems research. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 389-398).

Brunswik, E. (1940). Thing constancy as measured by correlation coefficients. *Psychological Review*, *47*, 69–78.

Brunswik, E. (1944). Distal focussing of perception: Size constancy in a representative sample of situations. *Psychological Monographs*, *56*, 1-49.

Brunswik, E. (1952). The conceptual framework of psychology. In *International encyclopedia of unified science* (vol. 1, no. 10, pp. 656-760). Chicago: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkley: University of California Press.

Brunswik, E. (1957). Scope and aspects of the cognitive problem. In H. Gruber, K. R. Hammond, & R. Jessor (Eds.), *Contemporary approaches to cognition* (pp. 5-31). Cambridge, MA: Harvard University Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carter, S., Mankoff, J., Klemmer, S. R., & Matthews T. (2008). Exiting the cleanroom: On ecological validity and ubiquitous computing. *Human–Computer Interaction*, *23*(1), 47-99.

Castro, L. A., Favela, J., & Garcia-Pena, C. (2011). Naturalistic enactment to stimulate user experience for the evaluation of a mobile elderly care application. In *Proceedings of MobileHCI* (pp. 371-380).

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi-experimentation. In M. D. Dunnette & L. M. Hough (Eds.) *Handbook of industrial and organizational psychology*. Palo Alto, CA: Consulting Psychologists Press.

Dallin, G. W. (2005). *Control and treatment of hot-dip galvanized surfaces*. Paper presented at the 97th Meeting of the Galvanizers Association.

Dhami, M. K., Hertwig, R., & Hoffrage, R. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*(6), 959-988.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of CHI* (pp. 345-352).

Gray, W., & Salzman, M. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, *13*(3), 203-261.

Guerini, M., Strapparava, C., & Stock, O. (2012). Ecological evaluation of persuasive messages using Google AdWords. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 988-996).

Hammond, K. R. (1998a). Ecological validity: Then and now. *The Brunswik Society*. Retrieved from http://www.brunswik.org/notes/essay2.html

Hammond, K. R. (1998b). Representative design. *The Brunswik Society*. Retrieved from http://www.brunswik.org/notes/essay3.html

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, *19*(4), 319-349.

Holtzblatt, K., Wendell, J. B., & Wood, S. (2005). Rapid contextual design: A how-to guide to key techniques for user-centered design. San Francisco: Morgan Kaufmann.

ISO. (2010). ISO 9241-21: Ergonomics of human system interaction—part 210: Human-centered design for interactive systems. ISO F±DIS 9241-210:2010.

Kieffer, S., Batalas, N., & Markopoulos, P. (2014). Towards task analysis tool support. In *Proceedings of the Australian Conference on Human-Computer Interaction.*

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57-78.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, *23*(2), 209-237.

Mayhew, D. J. (1999). The usability engineering lifecycle. San Francisco: Morgan Kaufmann Publishers.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., & Vera, A. (2006). Breaking the fidelity barrier: An examination of our current characterization of prototypes and an example of a mixed-fidelity success. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1233-1242).

Nielsen, J. (2012). Usability 101: Introduction to usability. *Nielsen Norman Group.* Retrieved from https://www.nngroup.com/articles/usability-101-introduction-to-usability/

Nielsen, J. (2014). *Turn user goals into task scenarios for usability testing*. Retrieved from https://www.nngroup.com/articles/task-scenarios-usability-testing/

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 206-213).

Norman, D. A. (1983). Some observations on mental models. *Mental Models*, *7*(112), 7-14.

Norman, D. A. (1984). Stages and levels in human-machine interaction. *International Journal of Man-Machine Studies*, *21*(4), 365-375.

Sangiorgi, U. B., & Vanderdonckt, J. (2012). GAMBIT: Addressing multi-platform collaborative sketching with html5. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (pp. 257-262).

Sefelin, R., Tscheligi, M., & Giller, V. (2003). Paper prototyping—what is it good for? A comparison of paper- and computer-based low-fidelity prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.*

Shackel, B. (1991). Usability—context, framework, definition, design and evaluation. In B. Shackel & S. J. Richardson (Eds.), *Human factors for informatics usability* (pp. 21–37). Cambridge, UK: Cambridge University Press.

Sonderegger, A., & Sauer, J. (2010). The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied Ergonomics*, *41*(3), 403-410.

Stoffregen, T. A., Bardy, B. G., Smart, L. J., & Pagulayan, R. J. (2003). On the nature and evaluation of fidelity in virtual environments. In L. J. Hettinger, & M. W. Haas (Eds.), *Virtual and adaptive environments: Applications, Implications and human performance issues* (pp. 111-128). Mahwah, NJ: Lawrence Erlbaum Associates.

Trochim, W., Donnelly, J. P., & Arora, K. (2015). *Research methods: The essential knowledge base*. Boston, MA: Cengage.

Tullis, T., & Albert, W. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Whitefield, A., Wilson, F., & Dowell, J. (1991). A framework for human factors evaluation. *Behaviour and Information Technology*, *10*(1), 65-79.

Zen, M., & Vanderdonckt, J. (2014). Towards an evaluation of graphical user interfaces aesthetics based on metrics. In *Proceedings of the 8th International Conference on Research Challenges in Information Science.*

About the Authors

Suzanne Kieffer is professor at Université catholique de Louvain and Vrije Universiteit Brussel, Belgium. Her research and teaching interests are in Human-Computer Interaction, Information Visualization and User Experience (UX). Her prior works focus on the design and evaluation of multimodal interaction styles for the engineering of eHealth and mHealth applications. Passionate about UX, she has been working for and with users for over 15 years and provides consultancy services to help software organizations successfully define and implement UX strategy plans. She especially enjoys raising to the challenge in agile software development settings. She holds a PhD in Computer Science from Université Henri Poincaré, Nancy 1, France.

Transactions on Human – Computer Interaction