2015 48th Hawaii International Conference on System Sciences

ECOVAL: A Framework for Increasing the Ecological Validity in Usability Testing

Suzanne Kieffer, Ugo Braga Sangiorgi, Jean Vanderdonckt
Louvain School of Management, Université catholique de Louvain
Place des Doyens, 1 – BE-1348 Louvain-la-Neuve, Belgium
{suzanne.kieffer, ugo.sangiorgi, jean.vanderdonckt}@uclouvain.be

Abstract

The term ecological validity refers to the study of the measurable correspondence between experimental design and ecological setting (i.e., representative users performing real tasks in their natural environment). Therefore, it is important to anyone intending to generalize the conclusions from the particular circumstances of an experiment to wider ecological situations. However, there is no operational definition that enables to objectively quantify the ecological validity of an experimental design. The ECOVAL framework offers a formal starting point to bridge this gap by helping to both define and increase the ecological validity in usability testing. A methodology, a tool and guidelines to increase ecological validity are presented in an industrial case study. The return on investment of increasing ecological validity is examined in terms of user efficiencies and usability practice. Finally, recommendations to increase ecological validity are introduced, contributions discussed, and outlines for future work presented.

1. Introduction

Does the design of an experiment reflect real tasks performed in a real environment? In particular, how representative are the environment, the stimuli and the tasks experienced by experimental participants? How to maintain an adequate scientific control for variables without artificial situations that produce irrelevant conclusions? What correspondences can be made between an experimental design and its related natural setting? Such questions arise while designing, running and interpreting experiments and relate to some extent to ecological validity.

Ecological validity refers to the extent to which a study comprises "real-world" use of a system [6]. The ecological validity of a study is low if the experimental design differs from the situations, the stimuli and the tasks experienced in a natural setting. On the contrary, it is high if it takes place in the participants' natural setting or if participants do not know they are taking part in a research study.

Why is ecological validity important at all? Essentially, it is important to anyone intending to generalize the conclusions or apply the findings from the particular circumstances of an experiment to wider ecological situations in which the phenomenon being investigated would naturally occur [1]. Ecological validity goes beyond the concept of external validity, which concerns generalizing to particular – and across the type of – persons, settings and times [12]. The term ecological validity refers to the study of the measurable correspondence between experimental design and natural setting. Therefore, it enables a generalization beyond the bounds of particular experiments [1].

Why is ecological validity important for usability practitioners and researchers? Similarly to the previous argument, issues regarding generalization of results arise while designing, running and interpreting usability studies. The ubiquity of computers and the sophistication of telecommunication equipment have stretched the challenge for usability practitioners and researchers. This has extended the scope of *context of use*, as defined by Shackel in [24], to more users, tasks, tools and environments. Let us consider in particular the *environment* dimension which refers to the physical environment where the situation "user-system" takes place [24]: e.g., workplace, school, and home. If the experimental design has not reproduced the user's environment in a representative manner, it will most likely lead to uncertainty about whether or not the study reveals anything about what users actually do in their natural environment.

There is no doubt that ecological approaches have been valuable in Human-Computer Interaction (HCI) with field studies and *Wizard of Oz* simulations. Yet, there are practical barriers standing in the way of ecologically valid evaluations [6]: (1) sampling plans should put together proper sampling of participants, ambiguities and errors; (2) evaluations should run over large amounts time, people, places, and target both existing and unanticipated activities; (3) monitoring the use of the system should remain as unobtrusive as possible in order to not interfere with the evaluation's outcomes; (4) tools should make it easier to develop prototypes for evaluation, minimize the cost of running evaluation, and facilitate data collection and analysis.

1530-1605/15 $31.00 © 2015 IEEE
DOI 10.1109/HICSS.2015.61

Overcoming these challenges needs iteration which is expensive. Therefore, it often goes along with tradeoffs between realism, unobtrusiveness, data sparsity, ambiguity, and cost/time [6], or between the different types of validity [3]. Regarding the latter, a study may have been adequately carried out, i.e., the study was internally and statistically valid. However, the generalizability of the findings may be reduced, i.e., the study suffers from problems with ecological validity because the experimental plan does not sample variables from the user's environment [1, 9].

Another problem from a theoretical perspective is that there is no formal definition of ecological validity. It can refer to "the extent to which a study comprises real-world use of a system" [6], or to "the extent to which evaluation methods support realistic use in realistic settings" [7]. Studies are described as paying careful consideration to ecological validity if they are conducted with "real users, real work goals and (in) realistic settings" [19] or "in real, even controlled, scenarios" [13]. It may give the wrong impression that conducting fieldwork is sufficient to ensure that the study is ecologically valid. Although fieldwork follows an ecological approach per se, conducting fieldwork is not sufficient to meet ecological validity as it does not justify the correspondences between experimental designs and natural settings [1, 11, 14].

Moreover, as stated loud and clear in [17, 22], there are no general criteria for objectively measuring this validity, nor for capturing the critical attributes of the real-world environment that are worth being studied. As a result, usability practitioners and researchers are left with determining whether or not an experimental design is ecologically valid, and whether or not a dimension of the context of use is worth being studied.

What is missing is an operational definition of the concept of ecological validity in HCI that enables to objectively quantify experimental designs as more or less ecologically valid. A methodological approach has been adopted here in order to address this issue: (1) identify the dimensions that characterize ecological validity in usability studies; (2) define a framework to quantify ecological validity across these dimensions; and (3) extract the priority between these dimensions.

This paper is a preliminary step towards the design of ecologically valid experiments in usability studies. Its primary contribution is ECOVAL, a framework that is intended for both usability practice and research. On the one hand, ECOVAL provides practitioners with guidance to satisfy ecological validity requirements in usability testing, and offers them the possibility to be more explicit concerning the tradeoffs they made. On the other hand, ECOVAL is a starting point to define the dimensions involved in ecological validity towards increased ecological validity in usability testing.

The paper is organized as follows. The next section discusses how the concept of ecological validity is currently addressed in HCI and in psychology, which helps explain how the ECOVAL framework was built. We report an industrial case study where the concept of ecological validity was put into practice with Gambit. Firstly, we explain how to increase ecological validity in usability testing. Secondly, we demonstrate the return on investment (ROI) of increased ecological validity in terms of user efficiencies, user satisfaction, and usability practice. Thirdly, we study the relative importance of the physical device used during the evaluation. Finally, recommendations to increase ecological validity are introduced, contributions are discussed, and outlines for future work are presented.

2. Related work

How is the concept of ecological validity currently addressed in HCI? There is, at least to our knowledge, no clear and operational definition of ecological validity that enables to objectively quantify an experimental design as more or less ecologically valid. Consequently, the term ecological validity can tacitly mean many things. It often refers to fieldwork [3, 6] which follows an ecological approach per se. It also refers to real-life or naturalistic setting. For example, Castro et al. [7] refer to naturalistic conditions as "high levels of fidelity while performing the tasks in the evaluation", whereas Guerini et al. [13] refer to ecological studies as studies where "subjects are totally unaware of being tested, testing is performed during their natural activity on the web". In addition, thanks to established usability evaluation methods such as Wizard of Oz [6] or relatively new concepts such as living labs [8], ecological validity has become rather popular over the last decade. A call for ecological validity begins to spread [6, 8, 18].

However, as highlighted by Carter et al. [6] and confirmed more recently by MacDonald and Atwood [18], there is a lack of published work and therefore a lack of understanding about how field evaluation is conducted in the practitioner's world. How are the tradeoffs between the different types of validity made? How significant is their impact on the validity of the findings? Is the application of ecological validity feasible and cost-effective? To answer these questions would not only improve usability practice, but it would most likely broaden the scope of usability research.

So then, why has ecological validity been so little and so timidly addressed? Like Bernstein et al. [3], we believe that it is precisely because of the tradeoffs between the different types of validity that have to be made while conducting field experiments.

In particular, the goal of trying to maintain scientific control (i.e., internal validity) seems to conflict with the goal of sampling the variables from the user's environment (i.e. ecological validity) [22]. Papers in which the internal validity has been sacrificed at the expense of ecological validity are very likely to remain unpublished [3]. We also believe that the magnitude and the confusion surrounding ecological validity seems overwhelming to usability practitioners and researchers trying to address it. This is illustrated by discussing how the concept of ecological validity has been addressed in psychology.

Egon Brunswik first introduced the concept of ecological validity in the 1940's [4] and refined it while discussing the representativeness (external validity) of psychological experiments [5]. Ecological validity has been persistently debated since the late 1970's [14, 15], which seems to have paralyzed the attempts to define general criteria for objectively measuring this validity and for capturing the critical and relevant attributes of the real-world environment. Nevertheless, Schmukler [22] succeeded in identifying three dimensions of ecological validity: the *environment* in which the experiment takes places, the *stimuli* used in the experiment, and the *task* (or behavior) required of the experimental participants. These three dimensions should be as representative as possible of the participant's everyday setting.

3. ECOVAL

The ECOVAL framework builds on the combination of Schmukler's three dimensions of ecological validity and Shackel's components in human-machine systems (i.e., user, task, tool and environment) [24]. Schmukler's and Shackel's earlier works have been selected because they are rather long-established in their respective field. Schmukler's dimensions and Shackel's components (Figure 1) were combined into ECOVAL, a framework consisting of six operational dimensions for characterizing and quantifying ecological validity: signals, objects, test medium, user interface, task, and behavior (Figure 2).

The *signals* and *objects* dimensions both refer to the stimuli used and describe how the real world environment is captured in the experimental design. Signals characterize the sensory inputs representing the environment (e.g. sounds, smells, colors, lighting, heat, cold, dust, and pressure). Objects characterize the physical objects present in the environment (e.g., buildings, equipment, furniture, tools, and artifacts).

The *test medium* and *user interface* dimensions both refer to the computer system being tested.

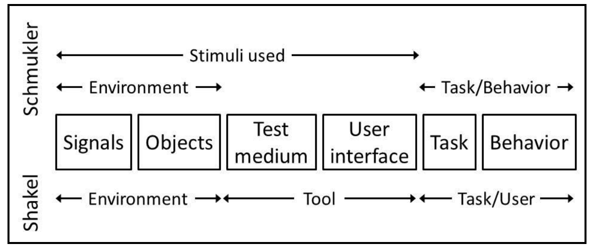

Figure 1. Schmukler-Shackel combination

Test medium characterizes the physical device used for interacting with the system, whereas user interface characterizes the software.

The *task* and *behavior* dimensions both refer to what is required of the participants. Task characterizes the experimental task performed by the participants. Behavior characterizes the behavior of the participants during the experiment.

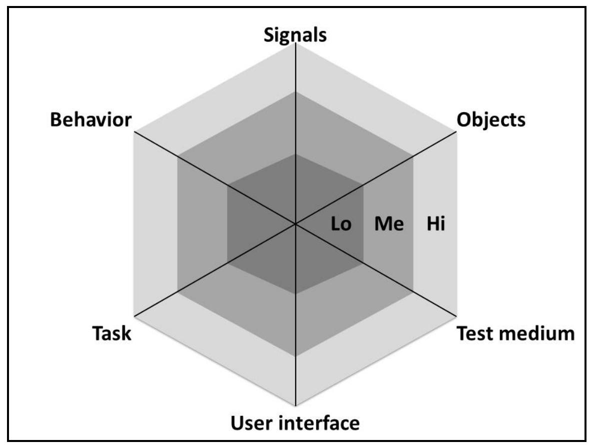

Figure 2. The six dimensions of ECOVAL

Other dimensions were deliberately omitted because they characterize other types of validity: the *experimenter* and *instruction* dimensions that both characterize the internal validity and the *user profile* dimension that characterizes the external validity.

The framework enables to quantify the ecological validity for each dimension separately, and across the six dimensions, so as to quantify the overall ecological validity of the experimental design. The ecological validity is quantified according to three degrees: low (Lo), medium (Me) and high (Hi). The degree corresponds to the level of similarity between the real world and what is captured from it and reproduced in the experimental design. Examples of low, medium and high ecological validity are presented in Table 1.

When appropriate, the notations Lo-, Me-, and Hi-ECOVAL will be used in the paper to refer to low, medium and high degrees of ecological validity.

Table 1. Examples of low, medium and high ecological validity for each dimension

	Low	Medium	High
Signals	No signals	Synthetized signals	Real signals (dust, noise, heat, pain, etc.)
Objects	No objects	Mock objects	Real objects
Test medium	Paper	Mock device Different device	Intended device
User interface	Video Storyboard	Prototype Mockup	Final interface
Task	Only verbalized	Mimicked and possibly verbalized	Real usage Real manipulation
Behavior	Only verbalized	Mimicked and possibly verbalized	Real actions (moving, talking, inspecting, etc.)

The degrees of ecological validity were quantified as follows. A low degree was assigned to artificial situations: when the phenomenon under investigation is "stripped" of its accessory, noisy, and messy characteristics [1]. Signals are artificial if the experiment takes place in laboratory conditions (see Table 1, "no signals"). Tasks are artificial when they differ from what is experienced in the real world, e.g. when the task is verbalized instead of being effectively performed (see Table 1, "only verbalized"). Let us consider the user interface dimension. Videos and storyboard were assigned a low degree because they are abstractions of what is experienced in the real world. On the other hand, a high degree was assigned to natural situations: when the phenomenon under investigation is kept intact or when it equals what is experienced by the participants in the real world.

Finally, a medium degree was assigned to the situations that fall in between. Whereas such an assignation seems obvious for signals, objects, test medium and user interface, it needs more clarification for task and behavior. How can the tasks and the behaviors be mimicked? Regarding the former, let us consider a paper-based prototype, i.e., the ecological validity is low for the test medium and medium for the user interface. In such a situation, participants are usually required to literally mimic what they would normally do with a real system by pointing or sliding with their finger on the paper prototype. Regarding the latter, participants can mimic the actions they usually undertake to complete a task such as interacting with a colleague, wearing special equipment, or moving.

The framework and the guidelines to use it were presented in Figure 2 and Table 1 respectively. In the next section, an industrial case study is reported. It demonstrates how to increase ecological validity in usability testing, the ROI of increasing the ecological validity and the relative importance of the physical device used during the evaluation.

4. Case study

The case study deals with usability testing carried out in the design phase of a mobile system developed to automate monitoring activities at the workplace.

It took place in a company whose core-business is hot-dip galvanizing (GA) to produce rust proof steel sheet for the automotive industry. Hot-dip GA consists of 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. It is operated on a continuous line: the sheet is continuously fed through a cleaner, an annealing furnace, and then into a molten zinc bath [10]. As presented in Figure 3, the production requires a minimum of five operators: entrance, bath, completion, exit, and a mobile operator who supports his coworkers and takes care of the furnace. The production is continuous. Consequently, four teams work in rotating shifts of eight hours: morning shift (A), afternoon shift (B), night shift (C), plus one team at rest. The workforce of operators totals 32 people.

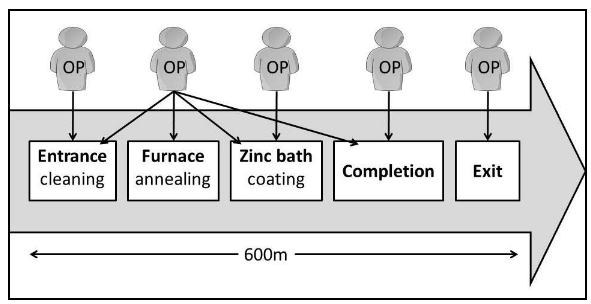

Figure 3. Work organization on the galvanizing line

The monitoring activity of the line is organized as follows. Operators are requested to fill-in with a pen a daily paper-based checklist. Each operator is assigned a specific list of items to check according to his posting on the GA line. Each checklist is structured as a 3-column table: the object and associated type of task (control, measure, or indicate) in the first column; the control value or requirement in the second column; and the task result in the third column. Once filled-in, the checklists are handed to the team leader for double-checking and validation.

The daily report totals about 300 items checked by operators and 100 items checked by technicians. Similar activities, e.g. equipment setting when changing the type of production, require similar monitoring activities. The system for the monitoring activity of operators could be reproduced, almost as is, to support other job functions and other activities.

4.1. Objective and procedure

In this context, our objective was to design, prototype, and test a mobile system for the automation of monitoring activities at the workplace.

We adopted a Usability Engineering (UE) approach and followed the methodology advocated by Mayhew [20] in order to address usability within the product development process. The following UE tasks were carried out: User Profile, Task Analysis, Usability Goal Setting, Screen Design Standards (SDS), and Detailed User Interface (DUI). A Style guide document was produced as deliverable.

Usability testing with representative users was carried out during both SDS and DUI. During SDS, paper mockups were evaluated with participants imagining that they were real product user interface. Once validated, the paper mockups were implemented as a prototype and evaluated again. During DUI, paper mockups were compared with a more realistic prototype of the same user interface.

The tool used to produce a more realistic prototype and to carry out usability evaluations is presented in the next section. Then, we give a practical explanation of how to increase the ecological validity in usability testing. Then, we examine the ROI of increasing the ecological validity in terms of user efficiencies, user satisfaction, and usability practice. Finally, we study the relative importance of the physical device used during evaluations.

4.2. Tool for prototyping and evaluation

Gambit [21] was chosen as the supporting tool for both prototyping and evaluation for two reasons.

Firstly, it facilitates the production of prototypes for evaluation and data collection. It enables usability practitioners to easily, quickly and cost-effectively produce interactive prototypes by simply interlinking specific areas of images (see Figure 4). Similar to wireframes, a Gambit prototype performs as a real product user interface by displaying images and enabling navigation paths between them. In addition, it can run on multiple devices: smartphone, tablet, laptop, monitor, TV, etc. Figure 5 shows Gambit running the monitoring report application for operators on a mobile device.

Secondly, it remains unobtrusive while monitoring the use of the prototype. It enables the data logging of events (i.e., screens that were looked through associated with date and time) and the live recording on the interaction map of the sequence of screens the user interacted with.

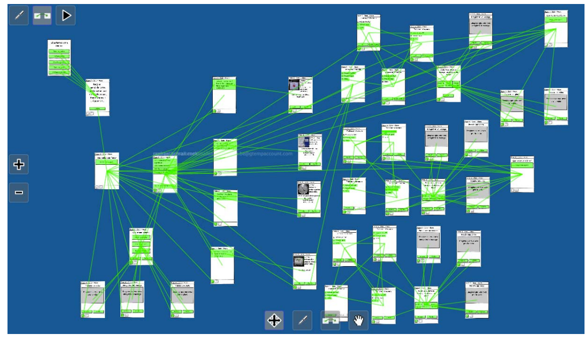

Figure 4. Images interlinked with Gambit to form an interaction map

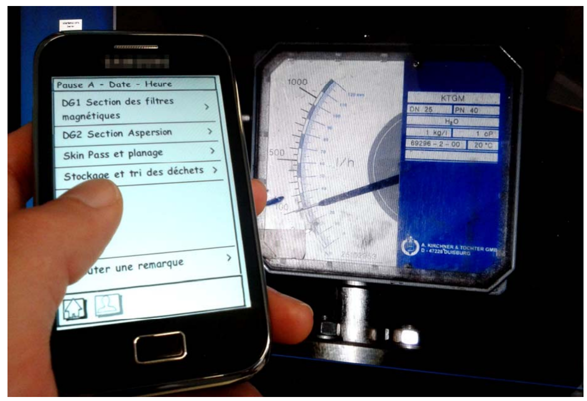

Figure 5. Gambit running an image-based prototype on a mobile device

4.3. How to increase the ecological validity in usability testing

Two rounds of usability testing with representative users were carried out during SDS with the aim of validating the presentation, the interaction with the application and the navigation paths.

The wireframing tool Balsamiq was used to produce the mockups of the user interface. The same Balsamiq material, i.e. the same screens and the same navigation paths, were used in both rounds. Consequently, the ecological validity of the user interface dimension was medium in both rounds. Paper mockups were used in the first round, whereas a Gambit prototype running on the intended device was used in the second round. Consequently, the ecological validity of the test medium dimension was low in the first round and high in the second round.

The SDS paper and Gambit experimental designs are both depicted in Figure 6 according to the six dimensions of ECOVAL (see the plain and the dash contours respectively).

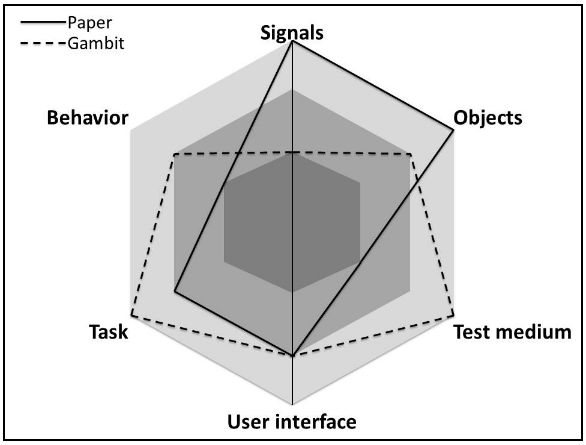

Figure 6. Ecological validity of Paper and Gambit prototype evaluations during SDS

Table 2. Description of the ECOVAL dimensions for Paper vs Gambit prototype evaluation

	Paper *First evaluation*	Gambit *Second evaluation*
Signals	Real (GA line) → High	Artificial (meeting room) → Low
Objects	Real (GA line) → High	Mock objects → Medium
Test medium	Paper → Low	Intended device → High
User interface	Balsamiq mockups → Medium	Balsamiq mockups → Medium
Task	Mimicked gestures → Medium	Real manipulation → High
Behavior	Only verbalized → Low	Real actions → High

As can be seen in Table 2, the paper mockup evaluation took place with real signals and real objects on the GA line. Participants were required to imagine that the paper prototype was a real product user interface, mimic the task and verbalize their behavior. Meanwhile, the Gambit prototype evaluation took place in a meeting room with artificial signals and mock objects. A mock object was the picture of a specific object targeted by an item on the checklist and displayed on a 21-inch monitor. As demonstrated in Figure 7, participants were required to walk from one monitor to another so as to perform four tasks T1 to T4. The participants were familiar with each object in the experimental setting. Consequently, they could easily recognize and check them. All participants successfully completed the four tasks: they all made a decision about the functioning of each mock object. Data related to user effectiveness were logged by Gambit as a pass or fail in selecting the region in the image where to enter their decision.

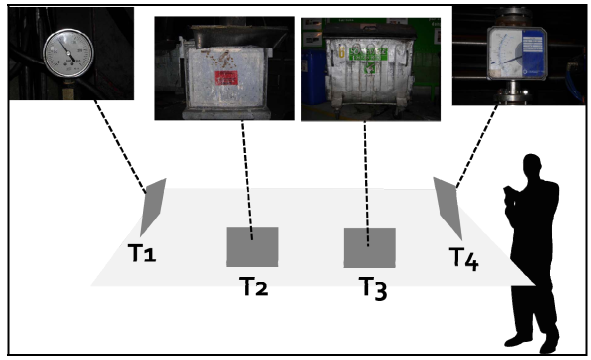

Figure 7. Experimental setting during the Gambit prototype evaluation

In order to maintain a defensible external validity in both evaluations, a stratified sampling of participants was made across postings (entrance, bath, furnace, completion, and team leader). Ten participants were recruited for the paper mockup evaluation and 18 for the Gambit prototype evaluation.

The paper mockup evaluation led to the validation of the proposed design as all participants expressed very positive judgments about the screens' content and look, the presentation of information and the clickable elements. They stated that they would successfully fill the monitoring report, and that the navigation was easy, fast and comfortable. Consequently, the Balsamiq mockups were implemented as a Gambit prototype and the second evaluation was conducted.

The ecological validity was higher in the Gambit prototype evaluation than in the paper evaluation. As demonstrated in Figure 6, the surface of the dash contour (Gambit prototype evaluation) is larger than that of the plain contour (paper mockup evaluation) with a ratio Paper-to-Gambit for ecological validity of 1:1.17. The benefits of increased ecological validity in usability testing are presented below.

4.4. Increased ecological validity and early detection and resolution of usability problems

Increasing the ecological validity had a significant impact on the early detection and the early resolution of usability problems. During the Gambit prototype evaluation, the first six participants consistently reported usability problems. They found "absurd" and "cumbersome" having to confirm their response for each item. They also reported difficulties navigating the user interface. Consequently, the Gambit prototype was modified accordingly before resuming the Gambit prototype evaluation with the remaining 12 participants. These changes increased user efficiencies and satisfaction.

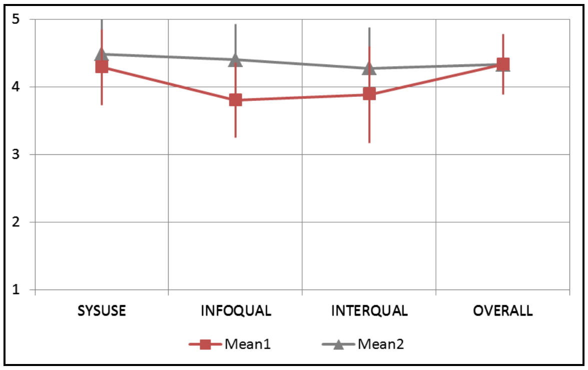

Figure 8. CSUQ stock chart - Mean1 (resp. 2) refers to mean scores of prototype1 (resp. 2).

The time in seconds and the number of navigated screens to complete each task were chosen as the measures of the users' efficiency and compared between the versions 1 and 2 of the Gambit prototype. Time and number of navigated screens were extracted from the Gambit log-file as the time elapsed and the number of screens looked through respectively to get a task done. A Computer System Usability Questionnaire [16] with a 5-point rating scale was used to compare the users' satisfaction between the prototypes 1 and 2.

User efficiencies were much better with prototype 2. It only took 20 seconds and 3.2 screens on average to complete the tasks with prototype 2, compared to 35 seconds and 8.6 screens on average with prototype 1. User satisfaction was also better with prototype 2. The all-item score with prototype 2 was 4.37 compared to 4.06 with prototype 1. Results are depicted in Figure 8.

The scores *system usefulness* (SYSUSE) and *overall satisfaction* (OVERALL) were both higher than 4.29 for prototype 1 and prototype 2. This indicates a high level of satisfaction amongst the participants regarding the usefulness of the system and the user experience with the system respectively. The score of *information quality* (INFOQUAL) was higher with prototype 2 (4.36) than with prototype 1 (3.81). This important difference indicates that the information was much easier to understand and better organized in prototype 2 than in prototype 1. Finally, the score of interface quality (INTERQUAL) was higher with prototype 2 (4.31) than with prototype 1 (3.89). This is also an excellent result which indicates that the participants preferred prototype 2 than prototype 1.

Better user efficiencies and better user satisfaction both indicate that the changes made to the prototype increased the usability of the system. This result is very consistent with the subjective general comments that were made during the debriefings. Participants who tested prototype 1 mainly focused on usability problems related to the presentation of information and the navigation paths.

Table 3. Cost in hours of Screen Design Standards

Screen Design Standards	Usability Engineer Hours	Operator Hours
Design	8	
Balsamiq screens	16	
Paper mockup evaluation	32	10
Analysis and documentation	8	
Gambit interaction map production	16	
Gambit prototype 1 evaluation	30	6
Balsamiq screens revision	4	
Gambit interaction map revision	8	
Gambit prototype 1 evaluation	60	12
Analysis and documentation	12	
TOTAL	**194**	**28**

Conversely, participants who tested prototype 2 did not even mention any such problems, and instead focused almost exclusively on their user experience with the system in the factory. Their feedback concerned operational organization (e.g., where and when to put the mobile device in charge), workday organization (e.g., how the system would impact a workday), or even predicted uses (e.g., how a similar system could support other tasks in the GA line).

Increasing the ecological validity mainly by using a Gambit prototype to conduct usability testing led to the early detection and resolution of usability problems. In the next section, the ROI of increasing the ecological validity is presented in terms of increased operators' productivity and decreased development costs.

4.5. ROI of increased ecological validity

In order to demonstrate the ROI of increasing the ecological validity, a cost-benefit analysis over SDS was carried out. The costs and benefits were calculated and compared. The costs of SDS are presented in Table 3: the Gambit evaluations cost 130 usability-engineer-hours and 18 operator-hours. The benefits were estimated in terms of increased productivity and decreased late design changes. Such categories of benefit were chosen because they align with business goals [2]. The business goals were stated as increasing the operators' productivity and reducing the design/testing/development costs. The benefits were calculated by following the methodology advocated in the chapter 3 of [2]. The relevant analysis parameters are presented in Table 4. The first six parameters were objectively measured either during task analysis or during usability testing. The hourly wage of usability engineers complies with the current market rate. As the company did not disclose any wages, the operator wage was estimated according to the current market rate in the automotive sector.

Table 4. Analysis parameters

Analysis parameters	Values
GA line running days per year	300
Tasks (items checked) per day	300
Time to complete task (seconds) with prototype 1	35
Time to complete task (seconds) with prototype 2	20
Gambit evaluation usability engineer hours	130
Gambit evaluation operators hours	18
Usability engineer fully loaded hourly wage	$160
Operator fully loaded hourly wage	$80
System lifetime (years)	5
Ratio of late-to-early design changes	4

Table 5. First year and lifetime benefits and costs

First year benefits	Values
Increased productivity	$30,000
Decreased late design changes	$66,720
Total	$96,720
Lifetime benefits	**Values**
Increased productivity 5 years	$150,000
Decreased late design changes	$66,720
Total	$216,720
Costs	**Values**
Gambit prototype evaluations	$22,240

The ratio of late-to-early design changes was predicted using conservative predictions [2]. Expected benefits and costs are presented in Table 5. The *increased productivity* was calculated as the operator-hours saved per year (375 hours) multiplied by the amount of tasks per year (900) multiplied by the operator wage. The *decreased late design changes* were calculated as the costs of late changes minus the costs of early changes. Consequently, a net benefit of $74,480 was expected in the first year, and a net benefit of $194,480 was expected over a system lifetime of five years.

4.6. Importance of the test medium dimension

Usability testing with representative users was also carried out during DUI, with the aim of studying the relative importance of the test medium dimension and its possible links with other ECOVAL dimensions. Similarly to SDS, the same Balsamiq material was implemented as both a paper and a Gambit prototype. Once again, participants were required to imagine that the paper prototype was a real product user interface, mimic the task and verbalize their behavior, whereas Gambit was running on the intended device. The shaded cells in Table 2 indicate the dimensions that remained similar in both SDS and DUI evaluations.

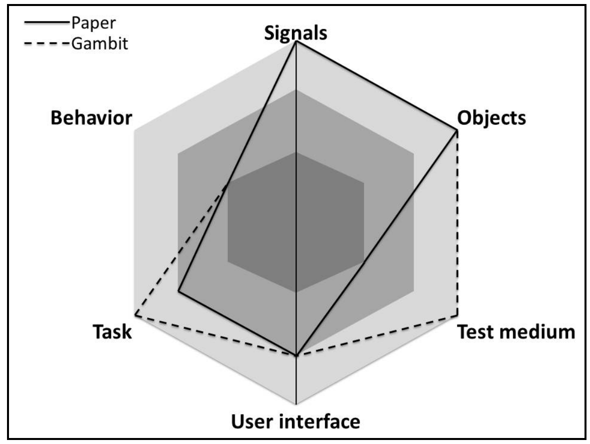

Figure 9. Ecological validity of Paper and Gambit prototype evaluations in DUI process

The DUI paper and Gambit experimental designs are both depicted in Figure 9 according to the ECOVAL dimensions. The study took place in the users' natural setting, where they usually are posted. The ratio Paper-to-Gambit for ecological validity was 1:1.43.

A stratified sampling of participants was made across postings (entrance, bath, furnace, completion, and team leader). Ten participants were recruited for this study. Each participant was required to perform the task with both test media. A Latin Square design was used to counterbalance the Paper vs Gambit conditions.

An interview was carried out immediately after the tasks to gather the participants' preferences between the two test media and their opinions about the study. All participants strongly expressed their preference for the Gambit prototype and argued for tool-supported evaluations, i.e. with the intended device. They stated that using a test medium with the "same look and feel" as the intended device made it easier for them to "project themselves into real use". They also stated that Gambit evaluations were "less intrusive" compared to paper-based evaluations where the experimenter "flips the sheets of paper as the interaction runs through". They also mentioned that they felt more "in charge" and "active" with Gambit, and more "passive" during paper-based evaluations. This result suggests that the choice of test medium has an important impact on the participants' motivation and commitment to participate to the study. This is consistent with previous findings where the choice of the test medium also impacts both the quality and the content of the feedbacks [23].

Finally, tool-supported evaluations, i.e. combining Hi-ECOVAL test medium with Me-ECOVAL user interface, lead to Hi-ECOVAL task. On the other hand, paper-based evaluations, i.e. combining Lo-ECOVAL test medium with Me-ECOVAL user interface, can only lead to Me-ECOVAL task.

5. Recommendations

The results of the paper suggest that the ROI of increased ecological validity lies in the early detection and resolution of usability problems, which increased employees' productivity and decreased late design changes. The results also suggest that users prefer high ecologically valid evaluations, especially compared to paper-based evaluations.

Primarily, we recommend increasing the ecological validity during usability testing with representative users as soon as possible in the Usability Engineering Lifecycle. According to the results of the case study, we strongly argue for experimental designs that run low fidelity prototypes (Me-ECOVAL user interface) on intended devices (Hi-ECOVAL test medium) and target real tasks (Hi-ECOVAL task).

Secondarily, we encourage the use of mock objects during the first iterations of design. Mock objects are simplified reconstructions of field events that produce data at least as reliable as that gathered in the field [6]. The advantage of using mock objects is twofold. On the one hand, they are cost-effective tools that produce reliable and usable data. On the other hand, they allow a defensible tradeoff between ecological validity and internal validity. The mock objects used during SDS certainly increased the ecological validity of the experimental task by providing participants with tangible instances of equipment to check. Additionally, the positioning of mock objects (Figure 7) led the participants to behave as expected in their natural setting, modulo the distance travelled. Meanwhile, having conducted the experiment with mock objects in a meeting room increased the study's internal validity by producing experimental conditions similar to those experienced in a usability lab: same instrumentation, same setting, and same signals (temperatures, lighting sounds, and stillness) for each participant [12].

Finally, we recommend increasing the ecological validity of the signals dimension when this intends to reinforce external validity, i.e. capture and reproduce environments in a representative manner in controlled experiments [1]. Lab situations with artificial signals seem appropriate for the testing of early design representations (e.g. SDS). Intermediate situations with synthetized signals seem to rather fit the testing of more detailed user interfaces (e.g. DUI). Uncontrolled situations with real signals seem to better apply after product fielding. Although the DUI evaluation took place with real signals, real world objects, real tasks, and the intended device, participants did not behave as expected. None of them went to actually check the equipment. Instead, they remained seated at the table. This strongly suggests that fieldwork does not necessarily lead to natural user behaviors.

6. Contributions

The paper addresses the problem of experimental studies' ecological validity that HCI practitioners and researchers alike have long struggled with. To what extend does the design of an experiment reflect real tasks performed by users in their real environment? This raises a further question related to external validity. Can the conclusions of a study be generalized to particular – and across the type of – persons, settings and times? These are important questions to anyone intending to increase ecological and external validities.

What is missing is a framework for measuring ecological validity and for analyzing the tradeoffs between the different types of validity made while conducting experimental research. The ECOVAL framework offers a formal starting point to bridge this gap. It consists of six dimensions: test medium, user interface, task, behavior, signals, and objects.

Another significant contribution of the paper lies is the demonstration of the operability and usefulness of ECOVAL in an industrial case study where practical guidelines and a supporting tool are presented. The ECOVAL dimensions seem robust, clear and reliable for manipulating ecological validity. Furthermore, increasing the ecological validity seems to increase the ROI of usability during design.

Yet, there are a number of other questions beyond the scope of this paper that need to be addressed in near future. What correspondences can be made between behaviors in an experiment and behaviors in the real environment? Moreover, how to transfer the natural/intended behaviors in an experimental context? A top-down ecological approach to tackle that relies on synthetic signals [1, 17]. Real signals are first recorded in the field. Then, participants' behavior, efficiency and satisfaction towards the synthetic versions of these signals are studied in the lab. Relevant signals in the case study are alarms, phone calls, and GA line stop.

Finally, this raises further questions. Is there a threshold in ecological validity that, once reached, ensures external validity? How do signals and behavior dimensions exactly impact external validity?

7. Conclusion

The ECOVAL framework offers a formal definition of ecological validity according to six dimensions: test medium, user interface, task, behavior, signals, and objects. The operability of ECOVAL and practical guidelines to increase ecological validity were depicted in a case study. The ROI of increasing the ecological validity in terms of users' productivity and decreased late design changes was also demonstrated.

8. Acknowledgments

The authors 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.

9. References

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

[2] Bias, R.G. and Mayhew, D.J. Cost-Justifying Usability, Second Edition: An Update for the Internet Age. Morgan Kaufmann, 2005.

[3] Bernstein, M.S., Ackerman, M.S., Chi, E.H., and Miller, R.C. The trouble with social computing systems research, CHI '11 Extended Abstracts, Vancouver, Canada, May 07–12, 2011.

[4] Brunswik, E. Organismic achievement and environmental probability. The Psychological Review, 50, 255–272, 1943.

[5] Brunswik, E. Perception and the representative design of psychological experiments (2nd edition). Berkley: University of California Press, 1956.

[6] Carter, S., Mankoff, J. Klemmer, S.R., and Matthews T. Exiting the cleanroom: On ecological validity and ubiquitous computing. HCI Journal, 2008.

[7] Castro, L.A., Favela, J., and Garcia-Pena, C. Naturalistic enactment to stimulate user experience for the evaluation of a mobile elderly care application. MobileHCI 2011, Stockholm, Sweden, Aug 30–Sep 2, 2011.

[8] Chi, E.D. A position paper on 'Living Laboratories': Rethinking ecological designs and experimentation in Human–Computer Interaction, Proceedings of CHI '09. San Diego, CA, USA, July 19–24, 2009.

[9] Cole, M., Hood, L. and McDermott, R. Ecological niche picking: Ecological invalidity as an axiom of experimental cognitive psychology, LCHC and ICHD, Rockefeller University, Los Angeles, CA, 1978.

[10] Dallin, G.W., Control and Treatment of Hot-Dip Galvanized Surfaces, Galvanizers Association Proceedings, 2005.

[11] Dunlosky, J., Bottiroli, S., and Hartwig, M. Sins committed in the name of ecological validity: A call for representative design in education science. Handbook of metacognition in education. New York: Routledge Ericsson, pp. 430–440, 2009.

[12] Gray, W. and Salzman, M. Damaged merchandise? A review of experiments that compare usability evaluation methods. Human-Computer Interaction, Volume 13, Issue 3, pp. 203–261, 1998.

[13] Guerini, M., Strapparava, C. and Stock, O. Ecological evaluation of persuasive messages using Google AdWords. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 988–996. Jeju, Republic of Korea, July 8–14, 2012.

[14] Hammond, K. R. (1998) Ecological validity: then and now. Retrieved in June 2014 from Brunswik Society website: http://www.brunswik.org/notes/essay2.html

[15] Kvavilashvili, L. and Ellis, J. Ecological validity and the real-life/laboratory controversy in memory research: A critical and historical review. History & Philosophy of Psychology, 6, pp. 59–80, 2004.

[16] Lewis, J.R. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. International Journal of Human-Computer Interaction, 7(1):57-78, 1995.

[17] Lewkowicz, D. J. The concept of ecological validity: What are its limitations and is it bad to be invalid? Infancy, 2(4), pp. 437–450, Lawrence Erlbaum Associates, 2001.

[18] MacDonald, C.M., and Atwood, M.E. Changing perspectives on evaluation in HCI: past, present, and future, CHI '13 Extended Abstracts on Human Factors in Computing Systems, Paris, France, April 27–May 02, 2013.

[19] Macleod, M. Performance measurement and ecological validity. Usability Evaluation in Industry, pp. 227–234, London: Taylor and Francis, 1996.

[20] Mayhew, D.J. The Usability Engineering Lifecycle. San Francisco: Morgan Kaufmann Publishers, 1999.

[21] Sangiorgi, U.B. and Vanderdonckt, J. GAMBIT: Addressing multi-platform collaborative sketching with html5. Proceedings of EICS '12, pp. 257–262. Copenhagen, Denmark, June 25–28, 2012.

[22] Schmukler, M.A. What is ecological validity? A dimensional analysis. Infancy, 2(4), pp. 419–436, Lawrence Erlbaum Associates, 2001.

[23] Sefelin, R., Tscheligi, M., and Giller, V. Paper Prototyping - What is it good for? A Comparison of Paper- and Computer-based Low-fidelity Prototyping. Proceedings of CHI'03. Ft. Lauderdale, Florida, USA, April 5–10, 2003.

[24] Shackel, B. Usability– Context, framework, definition, design and evaluation. In (Eds.), Shackel, B. and Richardson, S.J., Human Factors for Informatics Usability, pp. 21–37, Cambridge University Press, 1991.