

Oulasvirta, A. (in press). Field experiments in HCI: Promises and challenges. In P. Saariluoma, H. Isomaki (Eds.), *Future Interaction Design II*. Springer.

**This is the author's copy of the article.
Please consult the book for citation**

Field Experiments in HCI: Promises and Challenges

Antti Oulasvirta

Helsinki Institute for Information Technology [HIIT],
Helsinki University of Technology and University of Helsinki
Finland
antti.oulasvirta@hiit.fi

Abstract. Experimental methods have been under criticism since the advent of mobile and ubiquitous technologies, due to clear limitations in their suitability for studies in the field. However, the laboratory paradigm cannot be directly transferred to field conditions because of its strict notions of experimentation. This chapter examines the theory of *quasi-experimentation* as an alternative conceptualization of causality, control, and validity. Several threats to experimental validity in field experiments in HCI are discussed. These concerns must be addressed at all levels of experimentation, from the design and execution of a field experiment to analysis of data. Noteworthy also are new technical solutions that have enabled high fidelity data collection and that generally support endeavors in ensuring validity. If field experimentation is to become the de facto standard of research in human-computer interaction, the methodological core and technical tools must be developed in concert.

1 Introduction

Reason must approach nature in order to be taught by it. It must not, however, do so in the character of a pupil who listens to everything that the teacher chooses to say, but of an appointed judge who compels the witness to answer questions which he has himself formulated.

Immanuel Kant, 1781, *Critique of Pure Reason* (p. xiii)

According to Hacking (1983) and Shadish, Cook, and Campbell (2002), experimental procedures have been invented multiple times in the history of science. They note that Leonardo da Vinci's experiments in the 16th century and Galileo's 1612 treatise on floating bodies are considered landmarks in the natural sciences. In medicine, experimental procedures were employed to evaluate smallpox inoculation in 1721, and in Captain James Lind's studies onboard his ship to discover a cure for scurvy in 1747, as well as in Semmelweis's 1847 trials to reduce hospital infections. In 1879, Charles Sanders Peirce utilized randomization to investigate the psychophysical question of "just noticeable differences" in weights, and Hermann Ebbinghaus published a set of rigorously controlled experiments on his own memory a few years later in 1885. Statistician Ronald Fisher carried out the first randomized trials in agriculture, publishing the first coherent account of the methodology in 1923. During the 20th century, experimentation consolidated its position as the sine qua non scientific method in many if not most empirical disciplines.

Human-computer interaction is no exception. The early students of human-computer interaction (HCI) were strongly influenced by experimental methods in psychology. Paul M. Fitts (1954), based in Ohio, experimented on target acquisition performance by varying selection conditions, later synthesizing the results in an information-theoretical framework that was dubbed the Fitts' law. Douglas Engelbart's team at Stanford (Engelbart & English,

1988), working in the 1960s, converged through a set of experiments to the conclusion that the computer mouse, which they invented, is an optimal input device for an office information system in comparison to the lightpen and the tablet. The work at Palo Alto Research Center by Stuart Card and colleagues that lead to the cognitive model GOMS (Goals, Operators, Methods, and Selection rules) was based on a combination of experimental work and computational modeling (Card, Moran, & Newell, 1983). Finally, the 1990s saw a shift toward experimentation also in evaluation when Jakob Nielsen's (1993, 1995) usability engineering methods gained ground among HCI practitioners.

The argument for experimental methods in HCI has been the same as in experimental research in general—increased power in disentangling causal relationships from mere incidental occurrences. Experiments help illuminate complex chains of causal links, help distinguish between the validity of competing explanatory theories, and reveal descriptive causal relationship between conditions. By the same token, experimental methods have been viewed as central in endeavors other than hypothesis-testing, particularly in evaluation of constructed artifacts.

Even a cursory reading of HCI literature reveals that the paradigm of experimentation has been and still is confined to the laboratory. Recently however, arguments have been put forward that advocate experimenting in the field. Consider the prototypical abstraction of HCI: a user trying to accomplish a task in a command–feedback loop that includes the computer interface. The quintessential analytical constituents of interaction have been (a) the user, (b) the task, and (c) the interactive system. For a researcher perceiving interaction in this manner, there is no real need for conducting field experiments. Compare that framework to use situations of some prototypical ubiquitous and mobile applications: tourists searching for sights in a city with a location-aware map, a group of rally spectators discussing and sharing videos in a group media space, schoolmates messaging via mobile devices and PCs, commuters reading comics and watching TV on their mobile devices, information workers checking email in the backseats of taxis, or joggers sharing music during exercise. Our intuitions tell that there may be causalities in these situations that cannot be staged or reproduced in the laboratory, such as the geographical plan of the city, the rally event, users' own homes and schools, the train, the taxi trip, or the activity of jogging. To the extent that those have a causal role in interaction, the tripartite model of HCI is imperfect and incomplete, as is the laboratory as a setting for experimenting.

This is by no means a new message for students of HCI. The field of Computer-Supported Cooperative Work grew out of similar frustration when it was realized how prominently people and organizational dispositions feature as causal factors in the use of information systems. Analogous arguments can be found in papers dealing with activity theory (Kuutti, 1996), distributed cognition (Hutchins, 1995), and theory of situated action (Suchman, 1987). The divorce of these areas from HCI was so heated that the methodological premises of “the old” HCI was rejected with the theoretical one. Controlled experiments still seem rare in these areas.

There are two sufficient conditions for preferring a field study: first, an interest toward a causal agent that operates *external* to the human-computer loop and/or a suspicion thereof, and second, a belief that the causal chain wherein that agent operates cannot be properly reproduced or staged in the laboratory. In other words, field experiments are required when phenomena do not fit in the laboratory or cannot be simply staged there in a convincing manner. It would be nonsensical to conduct a study of typing performance on a mobile device in the field, unless one was interested in the effects of, say, real-world multitasking or lighting conditions. By contrast, valid evaluations of mobile maps can only be carried out in the field.

Against the backdrop of the success of the laboratory paradigm in HCI, and given its fundamental limitations in the context of newer technologies, it is surprising to note how rare field experiments are in present-day HCI. A metareview of mobile HCI research methods by Kjeldskov and Graham in 2003 summarized six concerns of researchers related to field studies:

- a) Time and/or personnel resources
- b) Skills and/or technological competence
- c) Control of experimental variables
- d) Expensive data collection
- e) The presence of researcher changing the phenomenon-of-interest
- f) Observations that do not generalize.

Kjeldskov and Graham conclude that the bias towards building systems limits the development of cumulative knowledge on mobile HCI. Four years later, the situation has not improved significantly. Some systematic field experiments have been built around the experience sampling method (ESM), but this way of data collection has been of limited applicability.

Now one can ask whether this state of affairs is due to insurmountable problems in the foundations of experimental methodology or due to our inability to discover solutions to the specific theoretical and practical barriers that field experimentation faces. History has shown that for a method to be widely adopted, sufficient levels of both theoretical and pragmatic maturity must be attained. The breakthrough of usability engineering practice, for example, was largely due to Nielsen's (1995) work in combining a theory of errors, a method for predicting the sufficient number of users, examples of experimental design, templates for measurements, and guidance for research instrumentation, such as "the usability lab" presented in Figure 1. The predictive modeling approach suggested in the bold manifesto of Card et al. (1983) never reached comparable popularity, despite the fact that their cognitive user modeling methodology was advanced, solved many pertinent problems, and it was theoretically coherent. As Hacking (1983) puts it, scientific breakthroughs are often based on a union of speculation and articulation, calculation, and experimentation.

The road leading to a sound basis for field experiments in HCI is undoubtedly rife with pitfalls. The present paper examines the theory of quasi-experimentation by Shadish et al. (2002) as an alternative to the prevailing laboratory experimentation paradigm. Only a very selective examination of their theory is possible here. The selection of the particular issues is based on this author's experiences from close to 20 field studies. The aims of the analysis are (a) to rethink what experimentation means, (b) to identify threats that are unique in field experiments in HCI, (c) to gather requirements for good experimental practice, and (d) to assess various tools that are available for researchers interested in embarking on field experimenting.



Figure 1. Two setups for experimentation in HCI. On the top, Jakob Nielsen’s (1995) laboratory setup at Sun Microsystems that worked as a model and baseline for many laboratories built around the world in the 1990s. On the bottom, one of the first published “mobile usability labs,” developed in joint effort between Nokia Research Center and HIIT (Roto et al., 2004).

2 Rethinking experiments as quasi-experiments

An Experiment, like every other event that takes place, is natural phenomenon; but in a Scientific Experiment the circumstances are so arranged that that the relations between particular set of phenomena may be studied to the best of advantage. In designing an experiment the agents and phenomena to be studied are marked off from all other and regarded as the Field of Investigation.

James Maxwell in 1876 (as cited in Galison, 1987, p. 24)

Causal relationship can be argued to exist if the cause preceded the effect, the cause was related to the effect, and we have no other plausible explanation for the effect other than the cause (Shadish et al., 2002). The “canon of discovery” proposes four general bases for inferring a causality from observations: (a) if observed phenomena have only one factor in common, (c) if observed phenomena are common except for one factor, (c) if a phenomenon changes systematically whenever a certain event takes place, or (d) if a phenomenon is

partially produced by known factors and there is only one factor that can produce the remaining part (Nagel, 1979). Consequently, the goal of experimentation is to create conditions, or “mark off a phenomenon,” so that a single factor can be attributed as the cause of an observed similarity, difference, change, or amount. If that can be achieved, there are statistical methods for distinguishing differences representing probable “true differences” from mere accidents.

Running an experiment in real-life conditions outside the laboratory, however, almost by definition undercuts experimental control and summons numerous threats to validity of scientific inference. It becomes increasingly difficult, at times even impossible, to eliminate alternative explanations for the treatment. The theory of *quasi-experiments* (Cook & Campbell, 1979) was founded upon the acceptance of the imperfection of field experiments as experiments—the degree of control is limited and should be treated as such. Having said this, it must be noted that some of these perturbations over which we have limited control can actually be of interest and should be treated not only as confounding factors. Experimentation should not be viewed as driven exclusively by hypothesis-testing. Two other motives for going into the field are (a) to learn about which real-world circumstances actually affect the phenomena at hand, and (b) to assess the robustness of that phenomena in those circumstances. The former goal calls for the ability of the experimenter to gather knowledge about those events and the latter for enough repetitions to be able to sift systematic interactive events from accidental ones. In both cases, the identification and mitigation of confounds is a central task of the experimenter.

The reward of experimenting in the field, improved realism, is achieved only by sacrificing ability to fully control events. The dual dimensions of experimentation—control versus realism—allow us to place types of experiments into an order of increasing realism and decreasing level of experimental control:

1. **Laboratory experiments.**
2. **Analogue experiments** are laboratory experiments that deploy simulations and emulations of real-world conditions to increase the generalizability of results. For example, the 1990s trend of decorating usability laboratories like living rooms can be conceived as an attempt to reproduce aspects of real use situations.
3. **Quasi-experiments** are experiments where an experimental intervention is carried out even while full control over potential causal events cannot be exerted. There can be systematic differences between experimental conditions that hamper the inference of causality to a single cause. For example, one can compare two notification mechanisms in PDAs in terms of perceived load and acceptance (Ho & Intille, 2005).
4. **Natural experiments** are “after the fact” quasi-experiments, where the variation of a causal agent has taken place naturally. An example is an experiment comparing two naturally formed, causally independent user groups in terms of some variable that differ between them—say, comparing adolescents to adults in terms of adoption, appropriation, and perception a mobile messaging service.

Common to all four types of experiments is that they rely on “variation in the treatment, posttreatment measures of outcomes, at least one unit on which observation is made, and a mechanism for inferring what the outcome would have been without treatment—the so-called counterfactual inference against which we infer that the treatment produced an effect that otherwise would not have occurred” (Shadish et al., 2002, p. xvii). An effect is thus the difference between what *did happen* and what *would have happened*. What makes the inference of that difference counterfactual is that the two outcomes cannot take place simultaneously. To mention an example, an intervention experiment calculates the

experimental effect by comparing dependent variables in two periods of time—for instance, Period A, use without the system, to Period B, use with the system. The experimenter then compares A and B to find out the impact of the system (for an example in mobile awareness research, see Oulasvirta, Petit, Raento, & Tiitta, 2007). The presumption is that nothing else than the treatment itself distinguishes the two outcomes.

2.1 Control and validity

The goal for a quasi-experimental scientist is to create approximations for the physically impossible counterfactuals. All experiments are limited and thus all results are limited. The central goal of a quasi-experimenter is to be aware of these limitations and address them properly in the design and analysis of experiments.

These approximations are created by implementing various forms of experimental control. The options for a field experimenter appear almost as plentiful as those of a laboratory researcher. One can consider direct intervention to change the environment, application, materials, or the task. Various forms of preselection concerning the user can be considered, as well as classic forms of inducement, such as changing instruction, feedback, or confederates' behavior. Nevertheless, not all controls can, in practice, be fully implemented in the field. In our field experiments regarding mobile Web (Oulasvirta, Tamminen, Roto, & Kuorelahti, 2005), for instance, it had been difficult to ask participants to relax and not focus on performance, particularly in hurried environments such as railway stations. In the terminology of Shadish et al., 2002, this is called interaction between setting and treatment. There are extraneous events that can produce random variation, interact with the to-be-manipulated variable systematically, and even prevent treatments from taking place.

According to the theory, there are four types of validity of concern to an experimenter (Cook & Campbell, 1979):

1. **Statistical conclusion validity:** Is there a relationship between the manipulated cause and observed effects?
2. **Internal validity:** Given that there is a relationship, is it plausibly causal from one operational variable to another?
3. **Construct validity of putative causes and effects:** Given that the relationship is plausibly causal, what are the particular cause and effect constructs involved in the relationship?
4. **External validity:** Given that there is probably a causal relationship from Construct A to Construct B, how generalizable is this relationship across persons, settings, and times?

A chain of logic exists in the order of these types. In order to question internal validity, one must have knowledge of statistical conclusion validity, and in order to question construct validity, one must have knowledge of internal validity and, finally, in order to question external validity, one must have knowledge of construct validity.

2.2 The challenge for HCI

Given this general approach, it is possible to start charting threats that are particularly severe in HCI and then examining the nature of possible solutions. With a list of validity concerns from Cook and Campbell (1979), one can approach potential problems in HCI quite systematically. Table 1 lists, quite extensively, threats that have conceivable relevance in HCI. The general impression is disheartening: The quirks and foibles of our present-day experimental practices place us detached from any ideals.

Locality of results is perhaps the classic issue that arises when discussing external validity in HCI. The problem is that nearly all experiments are highly local but have general aspirations. This problem arises from inevitable differences between the conditions that an experiment creates and those to which the results ought to generalize. This observation takes us to an important point that can be made from the perspective of quasi-experimentation: The debate contrasting laboratory and field experiments is built upon a false question. It is an apriorism to claim that laboratory experiments are “ecologically invalid” and field studies valid only because the former take place indoors and in circumstances controlled by the experiments.

Every experiment creates *boundary conditions* for certain phenomena to occur and those boundary conditions are either common or rare in the real world, independent of the researcher. Thus, ecological validity is a question that must be assessed based on understanding the causal factors affecting the phenomenon of interest. For example, a study of target selection performance with camera phones assumes certain selection distances, target sizes, illumination conditions, and so on, that may or may not affect performance “in the wild.” The influence of these factors is an empirical question in itself. However, if these factors are properly taken into account in a laboratory setting, it does indeed have ecological validity. By the same token, nothing guarantees that field experiments have ecological validity. For example, some researchers assume that walking is representative of mobile behavior and stage their field experiments so that users walk predefined routes.

Table 1. Potential Threats to Causal Inference in Field Experiments in HCI. Adopted and modified from Cook & Campbell (1979).

Threats	Examples
<p><i>1. Statistical conclusion validity</i></p> <p>a. Low statistical power</p> <p>b. Violated assumptions of statistical tests</p> <p>c. Fishing and the error rate problem</p>	<p>Random irrelevancies due to “noise” in real world conditions</p> <p>Incorrect tests, e.g., due to non-Gaussian distributions or unbalanced designs</p> <p>Many dependent variables, statistical tests not scaled accordingly</p>
<p><i>2. Internal validity</i></p> <p>a. History and maturation</p> <p>b. Testing</p> <p>c. The reliability of measures</p> <p>d. The reliability of treatment implementation</p> <p>e. Random irrelevancies in the experimental setting</p>	<p>User getting tired or equipment accuracy decreased due to component breakage</p> <p>Learning across trials</p> <p>Shaky videotape recordings missing events, the moderator unable to shadow the user</p> <p>Difficulties in properly instructing the subject when outdoors, e.g., due to noise</p> <p>User meets familiar people when doing a task on the street</p>
<p><i>3. Construct validity</i></p> <p>a. Inadequate preoperational explication</p> <p>b. Mono-operation bias</p> <p>c. Mono-method bias</p> <p>d. Instrumentation changes over time</p> <p>e. Mortality, differential drop-out rates</p> <p>f. Interactions with selection</p> <p>g. Ambiguity about the direction of causal inference</p>	<p>Defining user experience numerically, e.g., 1–7 in Likert scale</p> <p>All tests run by the same experimenter, whose personality may influence behavior</p> <p>Only one measure utilized</p> <p>Changes in equipment over trials, e.g., video cameras changing due to breakdown</p> <p>Drop-out rate higher in one condition, e.g., due to one interface variation being boring</p> <p>One user group benefiting from “a local history”, e.g., knowledge of the site of trial</p> <p>Did environment affect behavior or vice versa?</p>
<p><i>4. External validity</i></p> <p>a. Interaction of selection and treatment</p> <p>b. Interaction of setting and treatment</p> <p>c. Interaction of history and treatment</p> <p>d. Hypothesis guessing</p> <p>e. Evaluation apprehension</p> <p>f. Experimenter expectancies</p> <p>g. Confounding constructs and levels of constructs</p> <p>h. Interaction of different treatments</p> <p>i. Interaction of testing and treatment</p> <p>j. Restricted generalizability across constructs</p>	<p>Compared groups differ in terms of interest toward the piece of technology at hand</p> <p>Results obtained in one setting do not generalize to others</p> <p>Results obtained in particular days (e.g., holidays) do not hold</p> <p>Knowing that 2D map is being compared to 3D map may affect navigation behavior</p> <p>Trying to do one’s best in an expensive field test with nice moderators</p> <p>The moderator guides users unconsciously through habitual action, e.g., by walking ahead</p> <p>Selecting too extreme age groups for comparisons to understand the effect of age</p> <p>Claiming that results generalize to conditions where only one treatment is administered</p> <p>ESM questionnaires as a data collection method may trigger more “awareness” in users</p> <p>One construct telling a different tale than others, e.g., user experience not matching RTs</p>

To generalize, the external validity of a field experiment in HCI can be evaluated through analysis of:

- the nature of subject pool
- the nature of information and skills and social factors brought into the experiment
- the nature of using the technological application
- the nature of tasks and materials
- the nature of environment.

Even though threats to external validity often hijack debates about experimentation, there are other threats to validity that may be as important. Among the most pressing is the problem of sound statistical practices in field experiments in HCI. The questions that need solutions include how to deal with missing data, unbalanced designs, various typical confounds, non-Gaussian distributions, and so forth. These concern both statistical testing and experimental design, as the two should go hand in hand. Future work should provide tools, as well as, perhaps, templates for statistical analysis and encouraging examples in the form of successful cases.

Regarding internal validity, low fidelity of data, inadequate control, and confounds are the most arduous challenges. The example given in the next section illustrates how these can be addressed in the design of experiments. Construct validity, the representativeness of the manipulated cause, is often threatened by excess reliance on one data source in HCI, for example relying on a single questionnaire when examining users' acceptance of a system.

The final challenge discussed here is that knowledge of the effects of manipulable variables tells nothing about how and why those effects occur. The theory of quasi-experimentation is clear on the issue that experiments cater to causal descriptives but not causal explanations. As Pawson and Tilley (1997) argue, it is a mistake to treat field studies as "black boxes," to borrow terminology from software testing, that link manipulated variables to observable outcomes. A healthier approach is to construct them as "white boxes," where the researcher can "peek under the hood" to see which causalities produced the observed changes. This insists a shift in the mindset of an experimenter to include in the repertoire a more qualitative kind of analysis that targets the understanding of the chain of causal factors leading from the treatment to the observed outcomes.

3 Emerging tools and methods

HCI is not tied to any particular procedure of data collection; the field experimenter can use anything from interviews to observations to psychophysiology. One way to classify methods available to a field experimenter is to look at their temporal relationship with interaction.

Methods relying on subjective opinions collected in researcher-informant interactions, such as interviews and questionnaires and diaries, and user-produced materials in forms of photos and video clips, are perhaps the only means to gather information on the construction of meaning as it happened, and as such they are irreplaceable. On the negative side, from the perspective of studying interaction, they can only be administered after or before the moment of interaction that they are about. They rely on the user's account of what happened, and such accounts are known to be prone to biases, distortions, and omissions.

Third-person ethnographic observation methods, such as participant observation and shadowing, by contrast, allow for capturing the actual moments of interaction as they unfold. Consequently, the researcher can be better aware of the nature of data, particularly missing data. The physical presence of a researcher can, however, have some effect on the observed behavior and the nature of this effect is not well understood in HCI. Despite the positive

sides, this is perhaps the most expensive way to collect data. As a human observer is collecting the data, the reliability of the measures gives rise to a “human factors” question: How long and how accurately and systematically can the observer collect data and how accurate and reliable are the categorizations?

While these methods are sufficient for most purposes, there are tools emerging that may help overcome some of the associated problems. HCI researchers are in the fortunate position that the technologies they study can be adopted also as methodological tools for collecting data. Moreover, progress in science and technology often go hand in hand. A recent example can be seen in psychology, where the celebrated “Decade of the Brain” would not have been possible without the preceding advancements in applied physics that led to the production of a noninvasive and affordable brain imaging technology, the fMRI (functional Magnetic Resonance Imaging). The fMRI enabled access to the most intimate, unconscious workings of the human brain during psychological experimentation. Similarly, mobile applications themselves have enabled new ways of collecting data.

In the subsections that follow, two such tools—background logging and video cameras systems—are reviewed. These tools are similar in that they do not presume the presence of a researcher because a technical device replaces the researcher in the task of data collection. Both also enable capturing interaction as it happens, but with potentially lower costs than by human recorders.

3.1 Desirable qualities of data collection apparatuses for the field

General desirable qualities of a data collection apparatus for field experimenting in HCI include the following:

1. **Mobility:** The device moves with the moving user, capturing interaction reliably wherever and whenever it takes place, in both indoor and outdoor contexts of use;
2. **Capture of embodied interaction:** It captures both bodily (physical) and virtual components of command and feedback, as well as environmental events—understood broadly as encompassing social interactions—that may have an influence on those nuances of interaction that are under scrutiny.

These two desired qualities stem from the nature of the interaction with technical systems other than desktop computers, such as mobile and ubiquitous technologies. The three remaining qualities do not concern the phenomena of interest but rather the logic of quasi-experimentation. They are derived from the discussion of threats to experimental validity:

3. **Unobtrusiveness:** The system does not in itself bring about direct or indirect changes to interaction, particularly to those aspects that belong to the field of investigation;
4. **Support for multimethod approach:** It does not limit the researcher to one source of data but rather overcomes “the black box problem” by gathering indices of potential causes to observed events;
5. **Quality control:** It allows the experimenter to be aware of the reliability and fidelity of data captured both during and after the experiment, assisting in answering questions such as what caused missing data, from what situations are data gathered, how reliably the data corresponds to the actual situations they come from, and so forth.

Any single apparatus can only approximate these goals. Generally speaking, it is the goal of methodology developers to push the limit between the possible and the impossible. To illustrate two vastly different kinds of apparatuses that have emerged only recently, this section concludes by examining background logging and video-based observations. Both are powerful examples of applications of wireless technologies. Background logging on a

personal mobile device has several advantages: the logging devices moves with the user and is able to log everything that takes place on the device itself; it does not necessitate the presence of a researcher and can thus operate on a scale of time not easily viable by other means, it can be combined with other methods such as interviews and it allows real-time quality control. The second technology, a hybrid multivideo system has qualities that make it, in some respects, orthogonal to background logging. It is a special system that has to be installed on and worn by a user. It can capture more extensively the aspects of physical interaction and environment, it can be flexibly combined with all sorts of data gathering methods, and it is less prone to missing data and other threats than is background logging. The downside is that the system itself most probably has an effect on interaction and, because of this, its nature is limited to nonlongitudinal, “one-shot” experiments.

The theory of quasi-experimentation posits that there is nothing in a method or tool per se that would be ecologically invalid, unrealistic, or obtrusive. Rather, these qualities can only be evaluated in the context of an entire experiment where that method is deployed. It is in the concrete setting of an actual experiment where causal powers of the method itself manifest, setting boundary conditions for the validity of inferring causalities between the manipulated cause and the observed effects. It is these boundary conditions that create a gap between “the ought to” and “the actual.” Each particular method in turn bears idiosyncratic limitations on what comes to the reliable observation of a given phenomenon. Generally speaking, these dispositions include (a) the extent and content of recordings and, importantly, their relationship to the studied phenomenon, (b) random error and variation inherent in the measurement, and (c) the reliability of executing the method in an experiment. The two tools are evaluated from these perspectives.

3.2 Background logging in an intervention experiment

Smartphones are programmable mobile phones. The other main characteristics are their sensing capabilities, storage capacity, and built-in networking. Moreover, the phone’s status as a communication tool should not be forgotten: People carry phones around and use them in the management of social relationships. Another promising feature is its sensing ability. Current devices allow for automatic gathering of the following behavioral data: location, other devices in physical proximity, mobile communications, user’s commands and interaction with the device, calendar events, and device state (network coverage, battery level, charger status, alarm clock, silent/audible profile). Figure 2 illustrates what a log of such information looks like.

Scenario: Sending a short message to a contact to request a call back.

```
10h51m48s  Open Contact application
10h51m51s  Select the contact (Mika)
10h51m52s  Open the message composer and write the message
10h52m48s  Send the message
10h53m00s  Reception of delivery report
11h30m27s  Incoming call (Mika)
11h30m33s  Answering the call (recording starts)
11h33m59s  End of call

// Interaction log
20040623T105148  To foreground
20040623T105148  Showing contacts
20040623T105151  Items: [Antti 0 0 0]/Mika, loc: Exactum (1:00) 22 11 17//
                Renaud 0 0 0
20040623T105151  Items: Antti 0 0 0/[Mika, loc: Exactum (1:00) 22 11 17]//
                Renaud 0 0 0
20040623T105152  Sending SMS to: Mika, loc: Exactum (1:00) 22 11 17
20040623T105152  To background

// Context log
20040623T103507  profile:0 General (0 7 Off)
20040623T103629  area, cell, nw: 19000, 1952, RADIOLINJA
20040623T104620  devices: 0060579a6f70 [Janne] 0002eea07729 [Antti]
20040623T105148  UserActivity: active
20040623T105148  ActiveApp: [101fbad0] contextbook
20040623T105152  ActiveApp: [100058c5] mce
20040623T105248  SMS : sent msg #1053236 to Mika: "Please call me asap!"
20040623T105352  ActiveApp: [100056cf] ScreenSaver
20040623T105552  UserActivity: idle
20040623T113027  app event: STATUS: call
20040623T113027  app event: STATUS: call status 3
20040623T113027  ActiveApp: [100058b3] Phone
20040623T113033  UserActivity: active
20040623T113033  app event: STATUS: call status 4
20040623T113033  app event: STATUS: recording call
20040623T113359  app event: STATUS: recorded

// Communication log
20040623T105248  EVENT ID: 2268 CONTACT: -1 DESCRIPTION: Short message
                DIRECTION: Outgoing DURATION: 0 NUMBER: +123456789
                STATUS: Sent REMOTE: Mika
20040623T105300  EVENT ID: 2269 CONTACT: -1 DESCRIPTION: Short message
                DIRECTION: Incoming DURATION: 0 NUMBER: +123456789
                STATUS: Delivered REMOTE: Mika
20040623T113033  EVENT ID: 2270 CONTACT: -1 DESCRIPTION: Voice call
                DIRECTION: Incoming DURATION: 207 NUMBER: +123456789
                STATUS: REMOTE: Mika
```

Figure 2. An illustrative example from ContextLogger’s recordings (Raento, Oulasvirta, Petit, & Toivonen, 2005). The ContextLogger records a user’s interactions, contexts, and communications.

3.2.1 Example: An intervention experiment

To illustrate the use of a smartphone for quasi-experimentation, let us consider the studies reported in Oulasvirta, Petit et al. (2007). The starting point for that series of field experiments was the well-known fact that the success rate of mobile phone calls is relatively low.¹ Recently, the field of HCI has witnessed the emergence of “mobile awareness systems” to mediate real-time cues of other people’s current context and undertakings. Importantly, these awareness cues, such as another person’s current location or alarm profile, can be used to infer the presence, availability, responsiveness, or interruptibility of that other person. Some have expressed pessimism about whether such inferences would be systematically utilized by the users to reduce the number of failed or interruptive calls. Our aim was to test this idea in a field experiment.

The particular application studied was called ContextContacts (see Oulasvirta, Raento, & Tiitta, 2005, and Raento et al., 2005), which is an awareness software integrated into the phone book of a smartphone. It presents seven real-time “awareness cues” that are automatically, without user input, transferred and presented within a user group.

An A–B intervention methodology from clinical medicine and clinical psychology was utilized where a baseline of behavior was gathered in a period of time denoted by A, after

¹ In our studies, mirroring statistics gathered in Finland, only 45–75% (average by subject, 15 subjects, 3,969 total call attempts) of calls reached the intended receiver (Oulasvirta, Petit, Raento, & Tiitta, 2007).

which the technology (“the treatment”) was introduced for period B. In such a study, the effect or impact of the technology under study is defined as observed differences between the two periods. Because technology effects are often slow to emerge and depend on the interplay of social interaction and practice related factors, longitudinal studies are necessary. Three teenager groups participated in the study for a total period of time of 265 days. Throughout that time, ContextLogger was running in the background, recording all available information.

From the studies we gathered 370 megabytes of raw data, including short recordings from 667 calls, 56,000 movements, 10,000 activations of the phone, 560,000 interaction events with our applications, 29,000 records of nearby devices, and 5,000 instant messages.

Automatic logs of contextual data and interaction covered between 53% and 93% ($M = 73\%$, $SD = 14\%$) of the study period. Reasons for missing data include running out of battery, turning off the phone, as well as faults in the logger software. Yet, this data gathering method afforded a set of sophisticated high-resolution analyses, such as how often the cues were viewed on the phone, how this access was distributed between different locations such as school and home (as interpreted from location information of ContextLogger), how long the cues were looked at just before placing a call (and after an unsuccessful call), and how these cues referred to locations in the beginnings of phone calls (as manually coded from over 600 phone call recordings).

In the analysis phase, we separated the different variables, such as location, interaction and proximity, and loaded them into a relational database. Current values of variables could then be queried for any single point in time, allowing them to be correlated with calls, which were our main unit of analysis. The call recordings were used as focal points of interviews, and the recordings, together with interview data, were used to gain a qualitative understanding of the situations represented by the values of observed variables.

Concerning the impact of the awareness system on communication practices, the main findings of that study were as follows. One group exhibited an increase of 12 percentage points in the success rate of within-group phone calls during Period B, when the awareness application was used, and this turned out to be statistically significant. Both groups (to whom we could administer this analysis) looked at the phone book for a significantly longer time just before the phone call (the so-called pre-call delay measure, Figure 3) during the B period than the A period. The most frequent utilization of the cues was associated with the participants being mobile. Moreover, one user group learned to systematically relate location information at the beginning of their phone calls at a higher level of granularity in Period B than Period A. Objective data like this was in accord with the subjective, postevent interviews with the participants of the study.

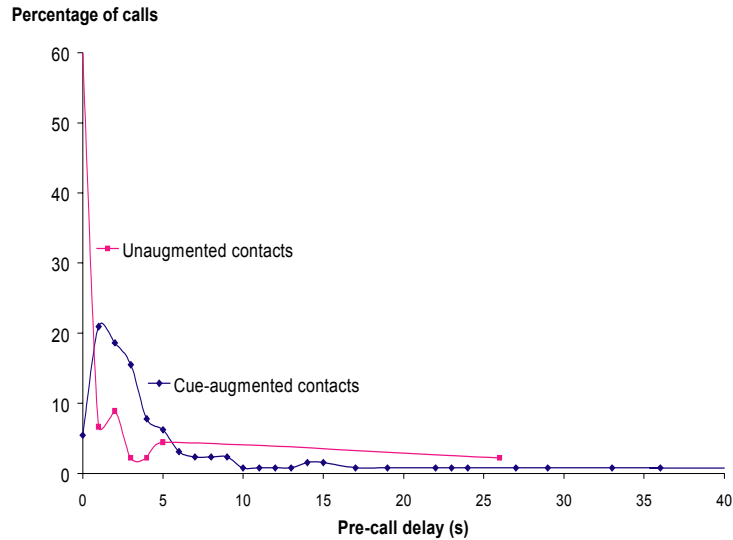


Figure 3. Distribution of frequencies of pre-call delays in the three trials. Those contacts for whom cues were available were viewed longer just before a call was placed to them.

3.2.2 Evaluation²

Counterfactual inference of causality in this particular study relied on comparisons between Periods A and B. Such a comparison is based on the ideal that use does not change between or during the periods. This assumption is of course problematic. For example, in one of the groups, the summer holiday took place during Period B, which lowered the frequency of use and changed the nature of communication. Furthermore, users learn and change over the period of the experiment, which should be taken into account in the intervention structure. For example, an A-B does not counter learning effects, and the A-B-A is still prone to accidental effects during Period B. Ideally, one would have several A-B pairs, but this in turn could disrupt the use itself if the period changed too often.

In addition to threats of counterfactual inference, there were other sources of threats to experimental validity stemming from the experimental procedure. One possible biasing factor was posed by the fact that we, that is, the researchers, paid the participants' phone bills, which most likely directed the group's communication to the smartphone and invited them to use the communications more regularly than they would have normally. Nevertheless, not all threats realized to the extent we were initially afraid of. The teenagers expressed no major technical or usability problems when changing from using their ordinary phones to the smartphones for the period of the study. The studied activity itself took place through the phone, so utilizing it as the data collection tool was natural. An alternative to smartphone-based logging would have been paper-based questionnaires or diaries asking the participant to mark how frequently they did something during a period of time.

More generally, smartphones can significantly reduce the costs required to record and log the mundane everyday activities of an informant and does not require an observer to be present in the activities. Improvements in ecological validity should be possible, since the automatic data collection can be done throughout the subject's everyday life and with minimal intrusion. The already available sensors can be used to infer many interesting aspects of an individual's everyday activities, such as movement at the macro level (based on GSM cell IDs), meetings and encounters with identifiable and unidentifiable people (Bluetooth presence), communications (phone calls and SMS), contents and use of contact book and

² This section is partly based on an article written together with Mika Raento and Nathan Eagle (to appear)

calendar, and audio scenes (microphone recording). The basic sensors can be supplemented, in principle, with more sophisticated ones, such as accelerometers and GPS for keeping track of movements at the micro level, physiological sensors for measuring emotions, and body-worn microphones for recording conversations. Analysis of the data can concentrate on individual events or more systematic patterns occurring over time. Depending on the population studied, the analysis can look at emergent patterns at the level of a social group, community, or geographical area. Thus, when applicable, smartphone-based data collection may augment self-report methods, offer in some cases a transition from self-report to observation, and extend the reach of experience-sampling, thus reducing the well-documented threats to validity of methods like diaries, interviews, and questionnaires.

Some of the data sources in phone-based logging are quite noisy. The Bluetooth-based detection of other subjects nearby is inherently stochastic. The absence of a signal in a Bluetooth scan cannot be used as proof that the person in question was not present. Noise per se is a threat only to statistical conclusion validity, given that the introduced noise is random. A more serious problem is caused by various inaccuracies. GSM-cell-based positioning, with city-and-district level tracking, may not give accurate enough location. It is, for example, not accurate enough to distinguish between home and the shop nearby, or an office and the lunch café. These inaccuracies can be systematic and thus should be accounted for in the analysis of data. On the positive side, foreseeable technological advances may help to overcome this problem. For example, we have explored with the possibility of augmenting location tracking with Bluetooth beacons set in appropriate locations, and one can entertain combinations with GPS-based as well.

When it comes to the content of data collected and its relation to communication behavior, the subject of study, some limitations are apparent. Studying communication patterns via the mobile phone will give strong insights into a subject's relationships, especially since both the occurrence of communication as well as the content of it can be collected. However, not all communication is through the phone, not even all technologically-mediated communication. E-mail and instant messaging can be used, even predominantly in some relationships. If comprehensive studies of communication are to be made, the e-mail and messaging data should be collected as well. It is quite easy to gather the e-mails sent and received by a subject, but detailed knowledge of the context in which a message was read or written may not be possible.

Another threat is posed by patterns in how people carry the phones. Although the phone is carried extensively by the user, it may be left behind by choice or accident. We have shown that detecting such situations is possible when the phone is forgotten for a significant period of time, but becomes considerably harder for short periods, for example, leaving the phone in the office when going for lunch. In general, it should not be assumed in the analysis of the results that all data gathered on the device corresponds to the activities of the user.

Studies conducted with the assistance of technology are of course susceptible to failures of technology. We have experienced faulty data connections, corrupted memory cards, crashing software, and broken phones. The most fragile link is often the data connection, which may be unavailable for days at a time due to failures of the phone software or lack of network coverage. Any study should take into account the possibility that remote real-time observation is not always possible. Even if remote data collection can be unreliable, so can be local collection. Software problems and hardware failure may result in losing locally stored data. In our experience, it is more reliable to gather data remotely, because the duration of a potential failure decreases significantly. If remote collection is not possible, data should be collected from the participants quite frequently, while accounting for the possibility data loss in the sample size and sampling strategy. The most extensive figures on the reliability of data

collection come from the Reality Mining study, where overall collection coverage averaged 85.3% (Eagle & Pentland, 2006).

While mobile phone technology is increasingly familiar to people in the developed world, not all users are comfortable or familiar with smartphones. Many mobile phone subscribers only use the most basic functionality and simple phones. Switching to a more complicated phone, or switching to a different manufacturer's phone, may scare some and will most certainly influence the way they use the device. If the subjects are not familiar with the smartphone, any measurements relying on phone use (communications, self-documentation, interaction logs) from the beginning of the study should be used with care. It is hard to give specific guidance on how long a "settling-down" period should be, but it may well be 1 to 2 weeks. It may be worthwhile to try to gauge how familiar the users have become with the device. At any rate, individual differences in the ability to use the phone pose a threat to validity, and thus should be addressed at the outset of research.

A problem in the HCI's practice of field evaluation has been the presumption that a single administration of the treatment is enough for evaluation, while in fact it does not allow for valid inference of the counterfactual. Excess reliance on "soft" baselines, the implicit or presumed baseline about the state of affairs as they are thought to be, is undoubtedly an unhealthy practice. The example above has illustrated, if anything, that there is a possibility to do proper comparisons, even in complex settings with emergent use phenomena. The A-B-A methodology illustrated above shows that it is possible to gather a concrete baseline on which to build the counterfactual inference.

3.3 Hybrid video system

A few video recording systems have been presented recently for the purpose of studying mobile use of information technology. Most of these systems are based on wearable cameras placed on the users, on the mobile device, or guided by the experimenter. For example, a system by Reichl, Froehlich, Baille, Schatz, & Dantcheva (2007) mounted a minicamera to a hat worn by the user, capturing the user's eyes, wirelessly sending data to a moderator who carried another camera and the recorder. Lyons and Starner (2001) presented a prototype where cameras and equipment were worn on user's body, in a vest. Applied Science Laboratories (n.d.) have recently presented a commercially available mobile eye camera. Google (Schusteritsch, Wei, & LaRosa, 2007) uses a system consisting of two cameras on the mobile phone for their studies.

Our earliest attempts in applying video recording systems are described in Roto et al. (2004) and Oulasvirta, Tamminen et al. (2005). Our later version, released in 2007 (Oulasvirta, Estlander, & Nurminen, in press), contains several improvements to operational capabilities, but also a more qualitative improvement: the ability to switch camera image in real-time between environmental cameras, for example when the user is moving in an office instrumented with cameras. Figure 4 presents the key components. Moreover, the whole setup can be carried on a belt, whereas the previous system required a backpack. Our other goals in developing this new version were (a) to extend the potential recording time to the length of a working day, (b) to increase the level of independent usage of the system without a moderator, and (c) to make the devices more compact, robust, and versatile.



Figure 4. Version 2 of Figure 1’s mobile video recording system. Some characteristics: (a) All noncamera equipment except cables fit on a belt and weigh less than 2 kilograms in total; (b) Environmental cameras can be switched to on the fly, based on signal strength; (c) All video and audio inputs are integrated on the fly to a four-video display; (d) The cameras are flexible, can be worn or attached to the mobile device, and can operate with cables or wirelessly; (e) There is an option for remotely triggered recording events initiated via Bluetooth; (f) Operational duration is up to 4 hours without battery change or other maintenance operations.

Special crafting was needed to develop further a research camera holder for a range of mobile devices. The resulting system consists of three parts:

1. Mobile/ “wearable” part of the equipment that has
 - one camera holder for fixing equipment to subject’s phone
 - one camera for imaging phone UI (to be connected with the holder)
 - one camera for imaging the subject’s face and behavior (to be connected with the holder)
 - a “Necklace-camera” for capturing roughly the same view that the subject sees
 - a wireless 2.4GHz video receiver
 - a video hub, the so-called video quad that gathers all of the video signals from several cameras to one recording device and provides adjustable voltage for the attached cameras.
 - a Video HDD mpeg4 recorder
 - three battery packs
 - a leather belt for carrying the devices
 - the necessary cables.
2. The semi-fixed part that can be either carried by the experimenter or placed to the environment consists of
 - surveillance cameras (3)

- wireless 2.4 GHz video senders (3)
 - 110-240 VAC to 12 VDC power adapters (3)
 - video staves and fixing equipment for surveillance cameras.
3. Tools for running and preparing the tests and maintaining the set-up
- battery rechargers
 - wireless receiver for setting up the environment cameras
 - travel cases for the equipment.

Figure 5 presents a diagram of the overall system architecture. To illustrate the use of the system, the following subsection describes a study utilizing it.

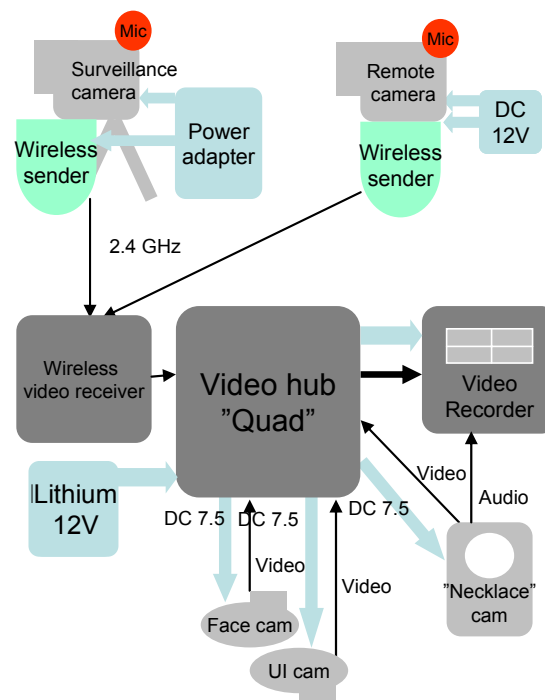


Figure 5. The system architecture of the system in Figure 4. The system consists of a “core” marked with gray, including the receiver, video hub, and recorder, as well as a batteries. All other parts are cameras that can be utilize either wireless or cable transfer.

3.3.1 Example: Comparing types of mobile maps³

To critically evaluate the system, we have conducted a field experiment following the logic of quasi-experimentation. Previous studies with the system would serve as examples as well (e.g., Oulasvirta, Tamminen et al., 2005), but the one reported here involves the most sophisticated experimental design and apparatus, and the most demanding mobile circumstances.

The starting point for this experiment was an interest in understanding how interaction with mobile maps differs between 2D and 3D representations (see Figure 6). Using a map is a process involving both mental and physical action, wayfinding, and movement (Darken & Sibert, 1996).

³ The study reported here has been designed and conducted with Sara Estlander and Antti Nurminen (Oulasvirta, Estlander, & Nurminen, in press).

In general, what makes mobile maps distinct from (typical) virtual environments (VEs)—such as virtual reality and desktop-based navigation systems—is that users are actually physically embedded in the world that the virtual model represents. When interacting with a physical environment (PE), shifting of gaze and movement of head and body are emphasized, whereas VEs are typically associated with decreased field of view and low fidelity of landmarks and nonvisual cues. This difference is crucial, as it means that users have to mentally construct the referential relationship between the virtual and physical information spaces. The hypothesis that these two representation types may differ stems from differences in interaction mechanisms: All movement requires maneuvering, performing a series of operations to achieve subgoals. Whereas maneuvering in a 2D view can be mapped to a few input controls in a relatively straightforward manner, movement in a 3D world cannot. In addition to 3D position, one needs to specify 3D orientation as well.

The particular interaction mechanisms of the studied are described in detail in Nurminen and Oulasvirta (in press). Here the focus is on the experiment and the role of the video recording system.



Figure 6. A 2D and 3D map view.

To study how users of mobile maps construct the referential relationship between points in the virtual space and in the surrounding physical space, a quasi-experiment was conducted in a city environment. The subjects ($N = 16$) conducted orientation tasks and navigation tasks. Three task types were used, and in each type the target was indicated on the map. First, in the *proximate mapping task*, the target was in view from the current position, and the task was to point to the target in the real world. During preparations before moving into the field, participants were instructed to turn to face in the direction of the target and point towards it with one hand. The instructions shown on the display of the mobile device before each of these tasks was the following: “Point to the target as quickly and accurately as possible.” Second, in the *remote orientation task*, the target was not in view from the current position, and the task was to point in the direction of the target in the real world. The instructions for these tasks were identical to those for the proximate mapping tasks. Third, in the *navigation tasks*, the target was not in view from the current position, and the task was to walk to the target. During preparations before moving into the field, participants were instructed to walk to the site of the target marker and stop on the pavement on the correct side of the street. The instructions shown on the display of the mobile device before each navigation task was the following: “Walk to the target as quickly and accurately as possible.” Altogether 24 search tasks were performed while moving a route of 2.4 kilometers in the old city center of Helsinki.

Taken together, there were two main independent variables: (a) Map type: 2D (traditional street map of Helsinki) versus 3D (the three-dimensional model of Helsinki); and (b) Type of task: orientation vs. navigation. In addition, several confounds were addressed in the experimental design:

- Because the misalignment between the target and the initial direction where the user faces affects how quickly the target will be shown, the facing direction was randomized for each subject and each task.
- To minimize the possibility of learning the areas of 2D map when using the 3D map, and vice versa, there were four loops (A, B, C and D) around at least one block, each with the starting point in the same area. Westerly and easterly loops were separated by a 300-meter distance. Half of the subjects performed 2D map tasks in the westerly loops (A and B) and 3D in the easterly loops (C and D), half 2D in the easterly loops and 3D in the westerly loops. Half of the subjects did loop A before B, half B before A, and within these two groups half did C before D and half D before C.
- To eliminate order effects, half of the subjects performed 2D tasks first, and half the 3D first.
- To eliminate effects of time of day, all experiments were conducted during daylight.
- To minimize effects of learning from one task site to another, each task's starting point was at least 50 meters from the previous.

The videocamera system, combined with other measures, allowed for a rich description of interaction. Five kinds of measures were employed:

1. Performance measures, for example task completion times, number of restarts, number of different types of key presses, and so forth.
2. Subjective workload ratings (here, NASA-TLX).
3. Interaction logs, analyzed with a custom-made visualization and replay software. The first interaction at the beginning of a task automatically started the logging.
4. Video data on users' interaction with the device and movement in the physical environment.
5. Complete verbal protocols during the task and retrospectively after its completion.

The output data, when combined with full transcriptions of verbal protocols, is an extremely rich source for analysis of the research question (see Figure 7). From the integrated data output, we manually coded the following:

- (User) walking: (a) User starts to walk to another position, "walking" referring to a series of more than four steps; (b) User ends the walk.
- Gaze-shifting behaviors: (a) User looks at the device; face is towards device. Supporting evidence involves fingers pressing keys in Camera 1 and changes in the playback of the interface; (b) User looks forward: face is forward or at a maximum of 45 degrees to the side relative to the body's sagittal axis and a maximum of 30 degrees up relative to the body's axial axis; (c) User looks up; face is at least 30 degrees up relative to the body's axial axis; (d) User looks left/right; Face is at least 45 degrees to the left/right relative to the body's sagittal axis, without moving feet.
- Bodily action: (a) User turns around by moving feet; (b) User turns the device; points the mobile device in a direction other than forward, at least 30 degrees.
- Interactive performance: (a) all key-presses and divided per key; (b) total distance traveled per task, in meters (camera in 3D, center crosshairs in 2D); (c) task completion times. Figure 8 illustrates navigation logs taken from the interactive device.



Figure 7. An example of output data from the mobile observation system of Figure 4.

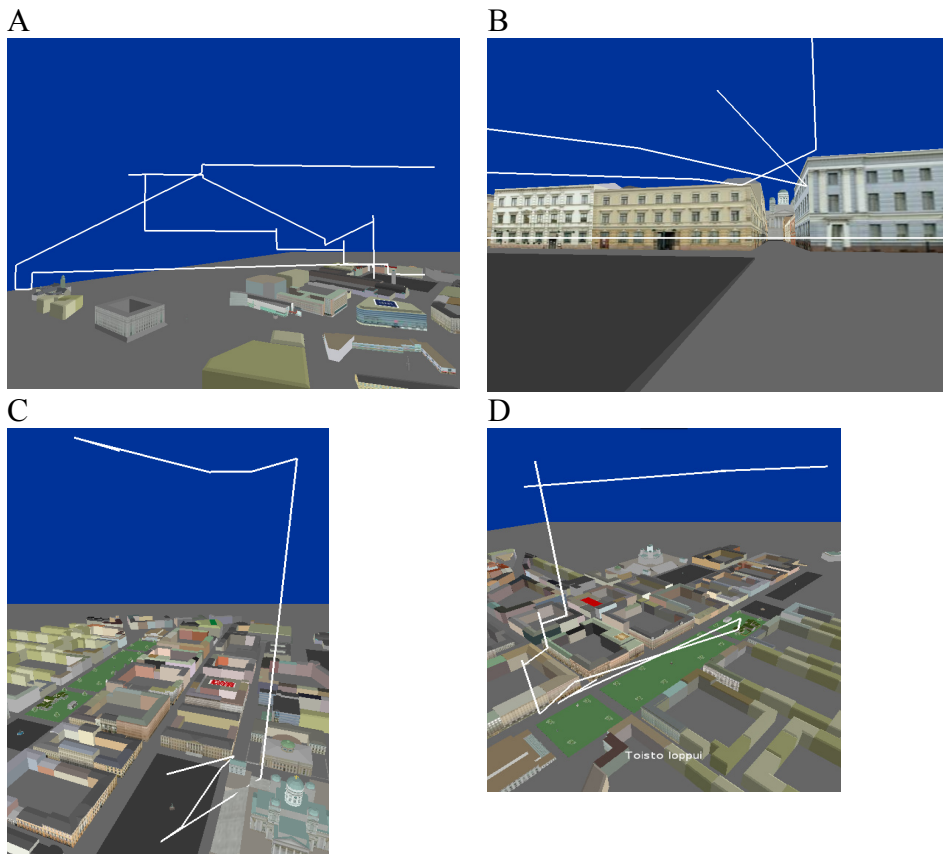


Figure 8. Visualizations of users' actions from the interactive logs: (A) Learning the model by exploring it from a top-down view; (B) Walking around in a city square and peeking around corners; (C) Diving to street-level; (D) Scanning, diving and scanning at multiple levels, and walking around at the street level. Adopted from a previous experiment reported in Oulasvirta, Nurminen, & Nivala (2007).

These codings and data were taken to answer a host of questions of interest when comparing 2D and 3D mobile maps, such as:

- How many times per task the subject looks at the device for more than 1½ seconds

- How many times per task the subject looks at the device for less than 1½ seconds
- Percentage of time per task subject looks at the device versus the environment
- Time from the start of task to looking away from device
- Time from the start of navigation task to starting walk towards the assumed target (walk to check the street sign does not count)
- Whether or not the subject walks at all in orientation tasks (proximate and remote)
- Whether or not the subject starts off in the wrong direction in navigation tasks
- Whether or not the subject chooses a nonoptimal route (not the shortest in terms of turning street corners) in navigation tasks
- Whether or not the subject first walks to the wrong side of the street in the navigation tasks.

▪

3.3.2 Evaluation

A critical assessment of the experiment centers on the question of how well it has been able to exclude alternative explanations. In the case of this experiment, the critical counterfactual inference is between the use of 2D and 3D. While the researchers took pains to eliminate and minimize interfering effects from 2D to 3D use, there is no practical way in a within-subject experiment to avoid all interference. Nevertheless, because of the full coding of both physical (bodily) and interactive behavior, we can analyze, *post hoc*, whether users utilizing 2D saw areas that they revisited in 3D to see if these had an effect on the dependent variables.

In addition to the question of whether our implementations of these countermeasures were effective, we have learned about two threats that are of more general importance for all field experiments employing video cameras. First is the question of how the video camera system itself affects the interaction. Several form factors can affect interaction with the studied system: The physical weight of the system may prevent or inhibit certain behaviors and cause fatigue, and the physical form may encumber movement and cause occlusion of objects in the visual field. While our system is based on minicameras weighing a few grams, it was clear nevertheless that it had a negative effect on how the mobile device was in use. The saliency of the camera system itself is a factor, as well. The camera system marks off the user as something extraordinary to other people, and can render the subject aware of receiving attention. Our design attempted to minimize the visibility of cameras, yet the two minicameras on a pole on the mobile device and the presence of a moderator following a person caught attention of passers-by. We have coded in the video transcriptions all such unexpected initiatives and can later exclude or include them in statistical analysis.

The second concerns the reliability of video recording as a measurement of subtle motor actions in field experiments. To be able to explain attained outcomes, our analysis of data was based on rigorous manual coding of events from the data, as reported above. Toward this end, we wanted to provide a full record of events leading to an observed outcome, for example, that a 2D map is better or worse than a 3D map in a certain task. In the coding, a problematic phenomenon surfaced: Certain environmental conditions and accidental events hampered the use of one or more of the minicameras, which made coding of certain variables impossible, or at least difficult. Direct sunlight in the face camera, the shutter adapting excessively to large contrasts in camera image, the necklace camera being temporarily obstructed by clothes, random compression artifacts, the experimenter-shot camera forgotten, and rain effectively preventing coding of some of the variables, particularly when the situation affected several cameras at the same times, thus disallowing the use of redundancies across the image sources. According to our analysis, these effects were primarily random and it was not likely that they impacted the experimental variables.

More worrisome were effects that were emphasized when the user was walking. The most significant effect concerns the learning of the experimenter to use the camera in a way that captures the bodily posture and gaze direction even when the user is walking. This requires walking at the same speed and at the same distance one or two steps behind the user to her left or right. When the user is standing, this task is trivial, but when the user is walking, the experimenter has to take care in walking as well, avoiding fellow pedestrians, trying to match the pace of the subject, and so on, and this requires some skill of its own. Despite several hours of practice gathered when administering the trials, some extreme walks were not adequately recorded even toward the end of the experiment.

We are not yet sure how critical the effect was on the quality and extent of missing data, or how to deal with it in the statistical analysis. Nevertheless, this problem was not as accentuated as in our previous experiment, where we could only utilize one camera that was directed by the experimenter.

4 Conclusion

According to Jonathan Grudin (personal communication, April 30, 2007), “the conundrum of HCI” is that to a person with imagination almost anything is possible, yet hard limitations exist that limit the use of technology. The discipline of HCI is therefore destined to work on two fronts: construction of the possible and empirical investigation of the impossible. Empirical work in HCI can therefore be viewed to entail two intertwined and complementary modes of research:

1. basic studies that aim at producing understanding of phenomena and factors relevant in human-computer interaction, and
2. evaluative studies of constructed prototypes that aim at producing informative and actionable knowledge for “extra-scientific” developers and decision makers.

The common denominator underlying both modes of research has traditionally been the fact that actions and reactions between computers and humans are the focus of a scientist’s analysis. In the studies of operators, programmers, managers, secretaries, students, and office workers as users, situational aspects have played a minor role. Events extraneous to the desktop have been presumed or thought to bear only incidental or unsystematic effects, and so it is unnecessary to include them in explanatory frameworks.

During the recent years, since the advent of mobile devices and ubiquitous technologies, this position has become increasingly more untenable. Consider a surgeon orchestrating the operation of a medical team, remotely, through a telepresence system; a tourist trying to locate a museum from a mobile map; a driver, turning the wheel with one hand and simultaneously calling home with the other; a group of teenagers coordinating via SMS where to meet; an information worker trying to synchronize his PDA with his laptop between two meetings; or a spectator browsing a digital pamphlet to decide which event to go and see. All of these examples have in common the fact that situational factors and events have a causal role in the course of events.

Consequently, empirical work in HCI should be able to shift to outside-the-laboratory settings. This challenge has been acknowledged and deemed particularly problematic to experimental methods. It may seem a paradox to suggest a controlled experiment in circumstances that deny full control. Therefore, one has to rethink what is meant by experimentation.

Toward this end, the possibility of utilizing the theory of quasi-experimentation as an alternative approach has been considered within this paper. It calls for marrying the design of

experiments with statistical analysis so that both take into account (real) threats such as missing data, unbalanced designs, random error in measurements, difficulties in implementing treatments, and so on. Furthermore, it calls for critical practices in evaluating experiments and reviewing research papers, habitually assessing de facto threats such as low grain of measurements, systematic biases in recordings, and obtrusiveness of the experiment, among others.

On the hardware side, this paper has shown two potentialities for research equipment, both associated with different threats to validity. Awareness of these problems may help not only in the task of critique but also in the task of constructing better experiments. And indeed, many tasks remain to be done. We need parallel advances in hardware design for apparatuses that enable reliable collection of data, software development for more efficient fusion and visualization of data, statistical methods to deal with typical problems, and pioneers who provide showcases illustrating the approach. While the primary aim is not to improve cost-efficiency but experimental validity, a revolution in the adoption of quasi-experimental methods can only take place if accompanied by “off-the-shelf kits” conceived, packaged, and marketed as products that appeal as worthy investments.

Despite the challenges, vistas for quasi-experimentation appear promising. In one sense, empirical research in HCI was problematized during the last decade when new personal and ubiquitous technologies appeared and demanded a radical shift in methods for empirical investigation. In retrospect, field experimentation did not secure the position it could have had in the toolbox of researchers.

The present paper has put forward a proposition that field experiments should be reconceptualized as quasi-experiments. The weaker form of the new paradigm of experimentation involves mainly more rigorous ways to address various confounds to validity. The stronger implication involves the idea that a central part of fieldwork in HCI, that concerning the evaluation of prototypes, can be rethought, formulated, and analyzed from a quasi-experimental perspective.

5 Acknowledgements

The author wishes to express gratitude to all collaborators, particularly Mika Raento, Sara Estlander, and Tuomo Nyysönen. Parts of the text on background logging are based on a manuscript written with Mika Raento and Nathan Eagle. This research has been funded jointly by the FP6 EU project PASION (FP6-2004-IST-4-27654) and the Academy of Finland project ContextCues. The camera system described in Figure 5 has been developed in the PASION project.

References

- Applied Science Laboratories. (n.d.). Head Mounted Optics. Retrieved July 30, 2007, from <http://www.a-s-l.com/prod-head.htm>.
- Card, S., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. New York: Houghton Mifflin.
- Darken, R., & Sibert, J. (1996). Wayfinding strategies and behaviors in large virtual worlds. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems (CHI '96)*; pp. 142–149. New York: ACM Press.

- Eagle, N., & Pentland, A. (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10, 255–268.
- Engelbart, D. C., & English, W. K. (1988). A research center for augmenting human intellect. In I. Greif (Ed.), *Computer-supported cooperative work: A book of readings* (pp. 81–106). San Mateo, CA: Morgan Kaufmann.
- Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381–391.
- Galison, P. (1987). *How experiments end*. Chicago: University of Chicago Press.
- Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge, UK: Cambridge University Press.
- Ho, J., & Intille, S. (2005). Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '05; pp. 909–918). New York: ACM Press.
- Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
- Kant, I. (1781/1999). *Critique of Pure Reason*. Cambridge University Press.
- Kjeldskov J., & Graham, C. (2003). A review of mobile HCI research methods. In *Proceedings of Mobile HCI 2003* (MobileHCI '03; pp. 317–335). London, UK: Springer-Verlag.
- Kuutti, K. (1996). Activity theory as a potential framework for human-computer interaction research. In B. Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction* (pp. 17–44). Cambridge, MA: MIT Press.
- Lyons, K., & Starner, T. (2001). Mobile capture for wearable computer usability testing. In *Proceedings of IEEE International Symposium on Wearable Computing* (ISWC 2001; pp. 69–76). Los Alamitos, CA: IEEE Computer Society.
- Nagel, E. (1979). *The structure of science: Problems in the logic of scientific explanation*. Indianapolis, IN: Hackett Publishing.
- Nielsen, J. (1993). *Usability engineering*. London, UK: Academic Press.
- Nielsen, J. (1995). *Usability inspection methods*. New York: ACM Press.
- Nurminen, A., & Oulasvirta, A. (in press) Designing interactions for navigation in 3D mobile maps. In L. Meng & A. Zipf (Eds.), *Mobile maps*. Berlin, Germany: Springer-Verlag.
- Oulasvirta, A., Estlander, S., & Nurminen, A. (in press). Embodied interaction with a 3D versus 2D mobile map. *Personal and Ubiquitous Computing*.
- Oulasvirta, A., Nurminen, A., & Nivala, A. (2007). Interacting with 3D and 2D mobile maps: An exploratory study (Tech. Rep. No. 2007-1). Helsinki, Finland: Helsinki Institute for Information Technology.
- Oulasvirta, A., Petit, R., Raento, M., & Tiitta, S. (2007). Interpreting and acting on mobile awareness cues. *Human-Computer Interaction*, 22, 97–135.
- Oulasvirta, A., Raento, M., & Tiitta, S. (2005). ContextContacts: Re-designing SmartPhone's contact book to support mobile awareness and collaboration. In *Proceedings of Mobile HCI 2003* (MobileHCI '03; pp. 167–174). New York: ACM Press.
- Oulasvirta, A., Tamminen, S., Roto, V., & Kuorelahti, J. (2005). Interaction in 4-second bursts: The fragmented nature of attentional resources. In *Proceedings of the SIGCHI conference on human factors in computing systems* (CHI '05; pp. 919–928). New York: ACM Press.
- Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London, UK: Sage Publishing.
- Raento, M., Oulasvirta, A., & Eagle, N. (to appear) Smartphone: An emerging tool for social scientists. *Sociological Methods and Research*.
- Raento, M., Oulasvirta, A., Petit, R., & Toivonen, H. (2005). ContextPhone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing* 4, 51–59.
- Reichl, P., Froehlich, P., Baillie, L., Schatz, R., & Dantcheva, A. (2007). The LiLiPUT prototype: A wearable environment for user tests of mobile telecommunication applications. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07; pp. 1833–1838). New York: ACM Press.

- Roto, V., Oulasvirta, A., Haikarainen, T., Kuorelahti, J., Lehmuskallio, H., & Nyysönen, T. (2004). *Examining mobile phone use in the wild with quasi-experimentation* (Tech. Rep. No. 2004-1). Helsinki, Finland: Helsinki Institute for Information Technology.
- Schusteritsch, R., Wei, C. Y., & LaRosa, M. (2007). Towards the perfect infrastructure for usability testing on mobile devices. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07; pp. 1839–1844). New York: ACM Press.
- Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs*. Boston, MA: Houghton Mifflin.
- Suchman, L. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge, UK: Cambridge University Press.