Evaluation of Prototypes and the Problem of Possible Futures

Antti Salovaara[1,2], **Antti Oulasvirta**[3], **Giulio Jacucci**[1]
[1] Helsinki Institute for Information Technology HIIT, Dept. of Computer Science, University of Helsinki
[2] Department of Information and Service Economy, Aalto University
[3] Department of Communications and Networking, Aalto University

ABSTRACT
There is a blind spot in HCI's evaluation methodology: we rarely consider the implications of the fact that a prototype can never be fully evaluated in a study. A prototype under study exists firmly in the present world, in the circumstances created in the study, but its real context of use is a partially unknown future state of affairs. This present–future gap is implicit in any evaluation of prototypes, be they usability tests, controlled experiments, or field trials. A carelessly designed evaluation may inadvertently evaluate the wrong futures, contexts, or user groups, thereby leading to false conclusions and expensive design failures. The essay analyses evaluation methodology from this perspective, illuminating how to mitigate the present–future gap.

ACM Classification Keywords
H.5.2 Information interfaces and presentation (e.g., HCI): User interfaces—Evaluation/methodology

Author Keywords
Evaluation methodology; Field trials; Experiments; Usability studies; Prototypes; Future

INTRODUCTION
This essay discusses a discrepancy in HCI's methodology with possibly far-reaching consequences for the field. The discrepancy is rooted in Herbert Simon's well-known distinction between natural sciences and *sciences of the artificial* [64]. Simon observed that, while natural science aims for objective observer-independent findings, some sciences—the sciences of the artificial—examine the construction of human-created artefacts. In contrast to the natural sciences, sciences of the artificial are value-laden, as their purpose is to change the world. Simon's argument shaped the identity of engineering, computer science, design, and later HCI. As the widely used definition of HCI confirms, it has ties to both natural and design sciences: On one hand, HCI research strives to increase knowledge of the design of interactive computing systems; on the other, it studies 'phenomena surrounding them' [22, p. 5]. In the former connection, it deals with the construction of artefacts, while in the latter, it approaches its subject in the manner of an empirical scientist.

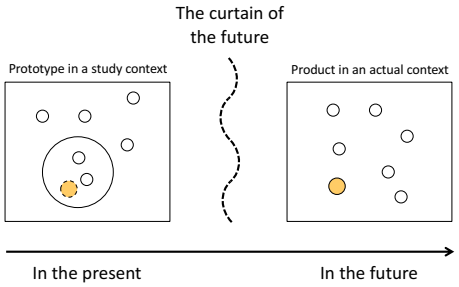

Figure 1. The present–future gap.

This discrepancy is essential in *evaluation methodology*. On one hand, HCI's interest in construction has been well understood [53]. Various methods exist for envisioning new concepts and for sketching and prototyping. At the same time, HCI's evaluation methodology is largely in intellectual debt to empirically oriented social and behavioural sciences that have not traditionally dealt with the construction of artefacts.

Our paper discusses *the present–future gap* (see Figure 1) that is particularly relevant in the evaluation of *prototypes*. Prototypes are, by definition, pieces of technology created for shaping and learning about possible futures and how they could be changed with technology. The construction of a prototype in itself, even without any empirical study, helps designers learn about the design problem [61]. However, empirical evaluation adds something further: ability to learn about its capacity to change the world. Hence, evaluations are about what the prototype *might* become. Although prototypes may embody research hypotheses that are future-oriented [75], only evaluations render them empirically researchable.

The future can be near or distant, and known with various levels of certainty. The evaluator needs to peer through this 'curtain of the future' but faces the dilemma that the future must be somehow enacted in the present if one is to draw conclusions about it, yet the future is inherently uncertain. In the present, a set of features is selected (represented by the white circles in Figure 1's left pane) for an evaluative study of a prototype (the dashed yellow circle), in an attempt to anticipate in the best possible way the future product (the solid yellow circle in the right-hand pane) and its future context. However, some differences between the two will always remain.

CHI 2017, May 06-11, 2017, Denver, CO, USA

DOI: http://dx.doi.org/10.1145/3025453.3025658

It is especially the boldest, most future-looking visions of HCI that need to pay attention to this issue. For an illustrative example, consider the first implementations of ubiquitous computing systems in classrooms, in the *Classroom 2000* project [1]. Here, we have the benefit of retrospection, since similar systems are now in use; however, in the 1990s, university students were not using phones, tablet computers, and laptops in the classroom. To learn about the impact of their new tablet-like note-taking application, the researchers evaluated it in circumstances that to some extent correspond with the way computers are used in classrooms in the 2010s. The studies did reveal some benefits that have since been documented, such as its usefulness for those students who do not normally take notes. However, because the prototypes were single-purpose devices, the Classroom 2000 work did not uncover the major riddle that mobile, networked computers pose for present-day education: distraction and avoidance of note-taking when materials are available online. Had early studies warned about this possible future, perhaps we would have fewer issues with ubiquitous technology in education.

This essay focuses on empirical evaluations of prototypes—usability tests, field trials, and experiments with novel technologies, which we call *evaluative studies* from here on—and the challenges in making valid claims about the prototype and its use in some future context. While the present–future gap is potentially critical for prototypes aimed at a distant future, the methodological considerations presented here are relevant also for 'near-future prototypes', albeit to a lesser degree. Near-future prototypes include new software iterations that may be evaluated in contexts whose features can be known with high certainty, such as rapid updates released via application markets. Far-future prototypes are ones such as those in the Classroom 2000 project, involving speculation and visionary thinking with increasing uncertainties.

Our focus is particularly on issues related to scientific validity. We discuss the construction of research questions and hypotheses, threats to validity, replicability, and separation between researchers and participants. This setting of scope allows us to address most of the empirical research published in the HCI field [13, 32]. However, it does not address all prominent research interests in HCI. For example, in the 'third paradigm' of HCI [20, 21, 74], an evaluation may aim at knowledge that is pluralistic, situated, and not necessarily valid in a traditional sense, to inspire new design ideas. Stakeholders may assume active roles that go beyond those of 'user' or 'participant'. However, as we argue, the present–future gap stretches our assumed 'scientific' notion of validity, because prototypes entail value-imbued assumptions about possible or desirable futures.

Our main point is that evaluators make a plethora of choices that influence the validity of the conclusions they can draw about the future their prototypes are supposed to create. These include the recruitment of participants; the physical, social, and computing settings wherein the study is conducted; the treatments to which the participants will be subjected; and the outcomes measured [63]. Indeed, as with the third paradigm, we will argue for explication of such values behind a study.

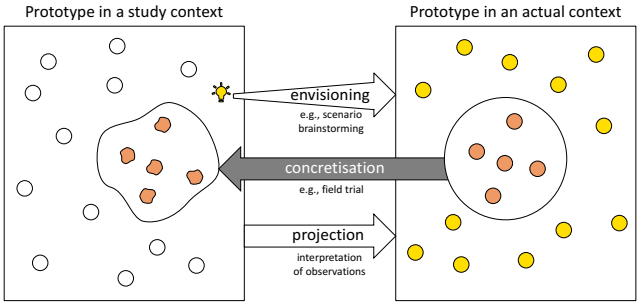

Figure 2. Three stages of prototype-based evaluative studies in HCI.

In the following, we first establish that HCI's *evaluation* methodology has not recognised challenges stemming from its inherent future-orientation. Secondly, we illuminate how common methodological choices affect a study's relationship with possible futures. We propose two dimensions of methodological choices, classifying them in line with their relationship to possible futures (control and staging). Finally, we discuss a change in mindset that is needed with regard to HCI's evaluations. Most of the argumentation we present is meant to open up issues for reflection and further research. We cannot claim to offer clear-cut solutions.

EVALUATION AS A TIME MACHINE

The first premise we want to establish is that any empirical study concerned with evaluation of a prototype has to deal with the present–future gap. While there are many methods for envisioning technologies, and sketching and prototyping them, the way we design our evaluative studies has not taken their future-orientedness into account. This is illustrated in Figure 2, which presents an overview of the methodological position of evaluative studies in HCI. Just as when we envision possible futures, when we set up studies we are readying a 'time machine' to peek behind the curtain of the future.

Let us consider the life cycle of a typical prototype and its evaluation. The first stage, *envisioning*, is the *only one of the stages that has been actively studied in HCI*. In this stage, teams work from their present-day understanding to envision what a new kind of technology (e.g., the next iteration of a product) could be and how it might be used. They enrich this future-oriented idea (depicted by the yellow light bulb in the figure's left pane) into some scenario for the future, represented by the yellow circles on the right in the figure. The challenges for this step, for coming up with plausible or compelling visions, are well recognised in HCI research [55]. Numerous concept design and brainstorming methods have been developed for this, along with explicitly future-oriented approaches such as design fiction [39] and use of trends as inspirational springboards [59]. User-involving methods, in turn, include participatory design [62], co-design [29], acting out [50, 54], and lead-user studies [35, 67, 68].

The second step—*concretisation*—is the primary topic of this paper. With this stage, evaluators make the vision empirically researchable in the present world. They need to focus on those features that are presumably important for the prototype's success or failure. Figure 2 depicts these in orange. Evaluators

then 'concretise' these features in the present. This involves decisions such as whom to recruit as participants or what tasks to assign to the subjects, in which setting to conduct the study, and how to indicate what should be done. But we look at them from the perspective of how such choices make some possible futures of the prototype empirically scrutinisable and observable. However, because of lack of resources, shortcomings in technical capability, and other factors, the concretisation is rarely perfect. This fact is illustrated with the features' distorted form in the left-hand pane of Figure 2. Later in the paper, we will discuss in more detail how the possibilities for concretisation differ between near- and far-future evaluative studies. This has a bearing on the differences between evaluations in research and company contexts.

The final stage is *projection*, referring to inferences that an evaluator draws about the prototype's future use, based on the empirical evidence gathered in the study in the present. This involves the challenge of making an inferential leap to the future while avoiding error-prone assumptions resulting from the present-day features. The projection should, in other words, be 'robust' in the face of alternative future trajectories of events. The 'validity' of the concretisation obviously affects the validity of subsequent projections.

Note that what we refer to as projection should not be confused with 'prediction'. The latter is aimed at reliable estimations of some future events in view of the information of the present. Projection, in contrast, has to do with *interpreting* findings from a study, obtained in the present time, with respect to what they suggest about the envisioned future. At its core, projection is about sense-making and need not be value-neutral. On the contrary, insights can be foregrounded selectively if they reveal interesting vistas for possible futures. If envisioning, concretisation, and projection are performed successfully, the evaluator may gain insight into the intended future worlds and the technology's role therein.

THE NON-ISSUE OF THE PRESENT–FUTURE GAP

The issue of the present–future gap has been recognised in HCI literature but overlooked in discussions of evaluation methodology. This necessarily cursory review argues that the root cause is that HCI has adopted its evaluation methods, in the most part, from fields that do not deal with artefacts.

To gain a better grasp of the landscape of HCI methodology, we need a concept to sift out those methods that are future-oriented from among those that are not. To this end, we define the *future-contingency of an evaluation method* as the extent to which choices in the implementation of the method are affected by consideration of the future. This definition is loose enough to prompt questioning of various methodological choices. Starting with the experimental method, we briefly review some of the main streams of methodology development in HCI, pointing to the conclusion that we have very few methods with explicit awareness of their future-contingency.

The Experimental Method

From its inception, HCI's evaluation methodology has been in intellectual debt to the behavioural sciences and, in particular, the experimental method [13, 32]. Take any textbook on HCI's methodology and one can find the experimental method as a key approach to evaluation. Its application in HCI puts emphasis on validity, reliability, and replicability [23]. The HCI field has borrowed techniques for mitigating nuisance factors and other threats to validity and for increasing the level of control. These were incorporated into an evaluation template by Nielsen [47]. A classic example of a widely known nuisance factor is the order effect (resulting from, for example, learning or fatigue), in which data collected at the beginning and near the end of an experiment are influenced differentially. To our knowledge, however, none of the nuisance factors addressed thus far in HCI research is explicitly future-contingent.

Another methodological element adopted in HCI is the 'divide and conquer' strategy. It calls for identifying components of some phenomenon and studying them one at a time by narrowing the experiment's research focus [42]. This approach, along with the concomitant fear of 'contrived' experiments, has been a target of serious discussion both in cognitive psychology [46] and, later, in HCI [3, 36, 73]. The strategy can be powerful, as the widely applied A/B testing method attests. It allows a product to be improved continuously with small decision steps evaluated in the wild and added to implementation as evaluative data dictate. Nonetheless, this strategy too was not developed with the present–future gap in mind.

Social Scientific and Socio-Technical Methods

The social sciences too have contributed significantly to HCI's evaluation methodology. The future is of interest in future studies, economics, and policy research, yet these fields of study seldom attend to human-made interactive artefacts.

Science and technology studies (STS) [7, 17, 40, 51] constitutes an exception with its interest in artefacts and their role in human life. Still, the way in which STS outputs have been applied in HCI work does not encompass future-contingent methods. While STS-informed studies in HCI (e.g., [11]) tend to project their findings to the future, their analyses focus on examining the present and past.

HCI has borrowed also from the ethnographic methods in anthropology (e.g., [25, 44, 66]). They are characterised by avoidance of control and intervention and by open-endedness [4]. These differences have been a starting point for several HCI-driven analyses of the relative merits and weaknesses of laboratory vs. field studies [33, 34, 57]. The researcher's passive role, while important for external validity, complicates evaluative studies wherein a prototype must be introduced to the participants' lives. As we will argue below, organising conditions in a study such that they resemble some future state of affairs requires some intervention.

Action research is an exception that refutes avoidance of interventions as an absolute. Instead, it deliberately aims for influencing and improving practices in the groups studied [71]. In that field too, a focus on human-made artefacts is seldom present. Action research's design-oriented adaptation, participatory design (PD), does take this approach, however, and is closely linked to HCI design practices. In PD, the future is

often enacted jointly with the participants to establish a common language with prospective users and elicit their contribution to the design [37, 62] through scenario-building and concept development. Hence, PD is a powerful envisioning method wherein the vision is determined by participants. The concretisation stage, with its idea of empirical evaluation of the future, is not present in PD, however.

Computer and Information Sciences

Computer science, unlike the behavioural and social sciences, deals mostly with human-made artefacts (algorithms, databases, networking, etc.). Its interest lies in improving the future via artefacts but it deals little with humans as users. Beta releases in the wild do stand as an exception. These evaluate a product's viability in the market and involve responsive development based on a user community's responses [45].

Information systems is another field with an interest in computer artefacts, placing special focus on managerial and organisational issues. With the exception of decision-support systems research [10, 48], work in this field almost never involves envisioning and construction of new systems.

Design Disciplines

Interaction design and design research overlap with HCI in their interest in human-created artefacts. The fields share many methods [72]. It is typical of design to put less emphasis on empirical evidence of outcomes and more on the process of creation. Consequently, although the design disciplines have recognised the present–future gap, they look at it mainly from the perspective of envisioning. The iterative prototyping method, for instance, may include imagining a prototype's use in some envisioned context, represented with storyboards and scenarios. While there are evaluation methods developed with the future in mind, they are developed primarily to inform design and not for developing defendable conclusions about interaction in future states of affairs. Theatre, bodystorming, and *in situ* scenario-development methods are good examples of means of concretising the future context [2, 28]. The focus is therefore on the 'recruitment of the future to design' [56] so as to inform envisioning. As noted, the epistemological stance in these fields is different. Instead of validity, reliability, and other concepts from scientific epistemology, design disciplines may focus on empathising with users and their culture (e.g., using cultural probes [8, 18] or 'provotypes' [9]), as well as on co-creation and participatory design alongside users [62]. This is in line with the humanistic research epistemology, wherein pluralism of values and knowledge, emphasis on users' agency, and elimination of the distinction between researchers and participants are prominent elements [20, 74].

However, some approaches within the design realm can be approached, alternatively, also with scientific epistemology. Research through design [19, 75, 76] and concept-driven design [65] frameworks emphasise design evaluations based on the prototypes' embedded theories [15] or their transformative capacity—the change that they have potential to induce in users' practices, activities, or experiences [75]. Here, technology serves as an instrument that foregrounds the phenomena of interest and makes them empirically researchable. Design researchers may evaluate their concepts also by arranging interventions and implementing them by applying experimental or ethnographic research methodologies. Their contribution to evaluative studies has been significant through the advocacy of speculative studies within HCI. They have shown the value of also studying undesirable futures (e.g., [5]) and other counterfactual settings on purpose.

HCI's Own Methodology Development

As the foregoing discussion shows, none of the 'mother disciplines' has addressed the present–future gap in such a way that HCI could directly adopt its work. What about HCI research itself? To sum up, instead of developing a systematic future-contingent methodology, its insights are scattered across isolated discoveries.

Usability evaluation methods were developed actively in the 1990s and are the standard approach in user-centred design. One clearly future-contingent practice is to set up laboratory studies in such a way that the conditions are realistic or feel so to the participants [58]. These set-ups can emulate a possible future. Another oft-used technique is sensitising scenarios (e.g., [14]): users are presented with scenarios of future use, to help them behave and think as if they were acting with a fully functional future product. In this way, the future is 'mimicked'. However, the validity of the concretisations and the projections is rarely problematised.

Field trials, as applied in HCI, borrow from ethnography and favour realism and openness over experimental validity. Typically (cf. [12]), a future use is examined by providing participants with a researcher-built prototype for use on their own. The prototype is introduced to participants' day-to-day life while researchers collect data on its appropriation [12, 26]. Prototypes are the primary means by which field trials address the challenges of future-orientedness: they render parts of the future concrete and observable.

UNPACKING THE FUTURE-CONTINGENCY OF METHODS

The second—and most important—point we make in this paper is that *an evaluator's choices should be understood from the perspective of their future-contingency*; that is, each choice in study design has potential to affect the presentation of the future in the evaluative study and, thereby, the evaluator's ability to draw valid conclusions.

To this end, we have analysed numerous HCI papers in order to pinpoint several techniques. Normally these are thought of as methodological choices from the perspective of considerations such as ecological validity. Here, we discuss these, summarised in Table 1, from the standpoint of the present–future gap. We do not claim that this taxonomy is comprehensive; its purpose is to illustrate the breadth of this matter. We proceed from the premise that all studies involve deliberately artificial (i.e., researcher-introduced) and sometimes even contrived changes in the study setting. This interventionist viewpoint is natural in this connection, since the prototype itself is always such an artificial change [15]. It does not exist in the present world outside its creator but might in the future. Accordingly, we align ourselves with Hutchinson et al., who

Technique*	How it can be applied	Contingencies with other *controlling* techniques**	Contingencies with other *staging* techniques**	Other contingencies
● Narrowing	Focusing on fewer aspects of the future setting	⇳ There are more nuisance variables to control	⇓ Fewer futuristic features need staging	Unnatural or 'wrong' futures; less generalisability to different futures
● Stabilising & removal	Making the features in the setting more similar across the participants	⇑ Work to find suitable physical research settings is trickier	⇑ Stabilised features must be made to seem natural	Fabricated observations may be produced
● Inhibition	Disabling, hiding, or blocking access to undesired features	⇑ May require blocking of other system functions in the prototype	⇑ Inhibitions may need 'camouflaging' as natural constraints	Fabricated results and less natural behaviours may result
● Gamification	Framing the technology and/or the study as a game	⇓ Through immersion in gameplay, there is less need for other control	⇳ (This depends on the nature of the game)	Suitable for limited purposes; game-design failure and/or fabricated findings may result
● Propping	Using physical props, mockup content, Wizard-of-Oz simulations, and human actors	⇳ Participants sticking to the desired range of actions may be an issue	⇓ This compensates for lack of other techniques	The cost may be too high or the *N* too small; concretising a 'wrong' future
● Setting selection & feature promotion	Employing a usage context or creating a UI design that increases desired behaviour	⇑ Inhibition and stabilising are necessary to keep participants in the setting	⇳ An engaging exaggeration decreases the need for propping, and an artificial one increases it	Observations (in an exaggerated context) may be setting-specific
● Repetition	Instructing or (subtly) guiding to repeat actions	⇑ More guidance must be prepared and enforced	⇑ More staging is needed to alleviate increased alienation	Observations might not be generalisable beyond task boundaries
● Recruiting	Selecting participants with specific profiles (e.g., lead users)	– (no effect)	⇑ Experts may require higher-fidelity prototypes	Undesired user-profile-specific behaviour patterns

* ● : The technique is primarily control-oriented. ● : The technique is primarily staging-oriented.
** ⇑: The technique typically *increases* the need for other techniques. ⇓: The technique *decreases* the need for other techniques. ⇳: The technique increases the need for some techniques and reduces the need for others.

Table 1. Techniques for bridging the present–future gap in an evaluation's concretisation stage.

'reject the strategy of introducing technology that only gathers "unbiased" ethnographic data' [26, p. 18].

The main distinction we make here is between *control-oriented* and *staging-oriented* techniques. These function in opposite ways in relation to the future. The main difference between these methodological components is depicted in Figure 3. Control is defined as attempts to restrict or inhibit present-day features that are not likely to be part of the intended future. Staging, in contrast, is an attempt to create or embellish futuristic features in the present. In general terms, the further away the future of interest is, the more staging and control are needed. An evaluative study may fail because it has too little or too much staging or control. With too little staging, the study in the present context does not sufficiently represent the intended future. Extensive staging, however, may be excessively costly and limit the number of participants or the duration of the study. Too little control may leave participants' actions overly heterogeneous, in which case they 'miss' the intended future. With too much control, in contrast, the study becomes too narrowly delimited, not allowing for participants' natural behaviour.

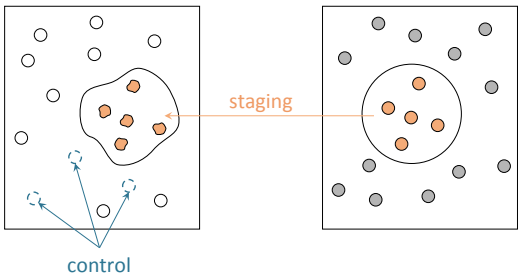

Figure 3. Staging and control components of evaluative study designs.

Control-Oriented Techniques

One way to consider the set-up of an evaluative study is to think of the evaluator's choices as *narrowing* the scope in its relation to the future (see Table 1). A study can never concretise the future to its full extent. This may be true even for near-future evaluations, as seen in iterative product development in a company context. Interactive systems are almost invariably interconnected within large ecosystems and infrastructures, and controlling the ecosystems' features is beyond the capacity of the evaluator. For example, it may be impossible to evaluate a webshop's online payment feature without integrating it with a full-fledged product catalogue and logistics. Such an evaluation may be forced to 'stick to the present' even if the new feature in itself would be radical. In contrast, systems that can 'stand alone' without dependences on ecologies allow more possibilities for narrowing and other methods. Such evaluations may be more commonplace in a research context.

Narrowing simplifies the study by limiting the number of futuristic features that need to be staged. However, it tends to increase the number of undesired contextual variables that then need to be controlled. This can lead to studying contrived or 'unnatural' futures: possible worlds that are crippled or lacking in crucial elements. Another threat is that of 'wrong futures': worlds that may be natural but do not represent the vision that the researchers had set out to study. Also, by making the future 'smaller', narrowing decreases generalisability of the findings. Hence, there is a risk that, as time passes, one finds little resemblance between the actual future and the one

that the study concretised. This leads to questioning whether the envisioned future was correctly specified in the study. On the other hand, narrowing the focus may afford better investigation of futures that are further from the present [42].

More generally, control is a familiar aspect of all experimentally oriented designs in HCI. It traditionally refers to actions by which a researcher unifies the conditions of the research context such that the resulting data show less bias and undesired variance. However, control has a future-oriented function too. As is noted above, it mitigates the conflicts between the present-day context and the possible future that the researcher desires to study. An example of a control technique used for such mitigation is removing all physical pens and paper from the research setting if the researcher is conducting a usability study of futuristic fully digital note-taking. We will now discuss this family of techniques more broadly, in the order presented in Table 1. We have visualised these, along with the above-described narrowing, in Figure 4.

Stabilising, *removal*, and *inhibition* limit, in different ways, the contextual features' natural variation. Firstly, stabilising allows nuisance factors to vary within tolerable limits. For example, experiments on mobile map navigation might be carried out in cloudy weather only, to stabilise the screen readability. Also, social settings can be stabilised, via use of actors who perform scripted actions in the same way for every participant [41]. Removal, in turn, eliminates nuisance factors from the research setting altogether. For example, walking on a treadmill may be used to simulate undisturbed walking in a city [6]. Finally, inhibition refers to noticeable limiting of contextual variation. For example, users in an evaluation of a futuristic communication tool may be instructed not to engage in communication via their familiar tools. In another example, a prototype might replace a Mac computer's taskbar (i.e., the 'Dock') with its own solution for launching programs [24]. Finally, usability evaluations make use of these methods extensively. They limit the context with the aid of scenarios with implicit or explicit statement of the tasks the user is asked to carry out. They also stabilise the interactions with assistance from digital mockup materials.

These three techniques are future-contingent, because they leave less room for participants to act freely. At worst, contrived behaviour emerges. Therefore, these approaches are safest when the techniques are used to limit interaction with only those features that do not exist in the envisioned future. Another safe use involves features that are irrelevant to the research question for the study. For instance, battery power can be safely stabilised in studies of hand-held devices if the research question pertains to new input techniques.

Gamification is sometimes used as a control technique, although it shares some features with the staging approach. Sometimes evaluators are concerned that participants will find evaluative studies uninteresting or socially awkward. This might lead to excessively self-conscious, unnatural, and disoriented behaviour. In these situations, the purpose of gamification has been to direct participants' attention to motivating features such as a captivating game. A user study may, for instance, involve an element of competition among the partic-

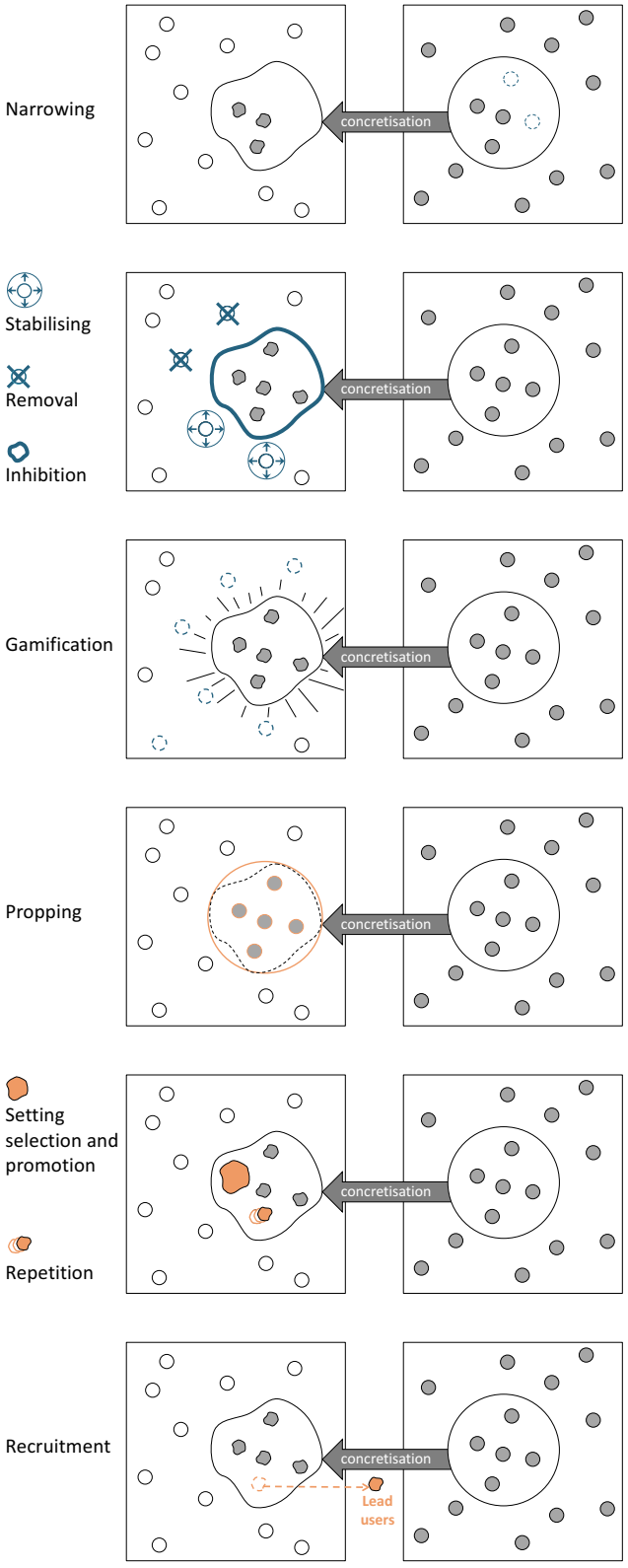

Figure 4. Techniques for controlling and staging futures in HCI's evaluative studies.

ipants, rewards for certain user actions, or target achievement levels that users are encouraged to exceed.

Game-like research designs, when the game is immersive and engaging, may then help participants ignore extraneous nuisance features (see Figure 4) that cause social awkwardness and other undesired effects. Therefore, other control techniques may be less necessary, since the excitement can help users stay focused on a smaller number of contextual features. However, the in-game elements need to be staged well if they are to sustain engagement. For example, a mobile augmented-reality (AR) map study by Morrison et al. [43] incorporated competition between teams. This helped to sustain interest in the study, which included a sequence of repeated tasks. Gamification cannot, however, be applied in every study, since not all activities can be made competitive without a change in their nature. It may, therefore, decrease generalisability: the actions carried out in the game are not necessarily the ones that the subject would perform outside the game.

Staging-Oriented Techniques

The control-oriented techniques discussed above, perhaps with the exception of gamification, are commonly seen in the experimental research although not typically discussed from the perspective of possible futures. Staging, in contrast, has been much more rarely discussed. The closest parallel in existing methods is the idea of analogue experiments [63]—for example, setting up a laboratory to resemble a living room. Staging is essential in evaluative studies since it generates an authentic feel for the study and includes all the relevant elements that make up the desired future setting, in the fidelity deemed necessary for valid inferences. Figure 3, above, illustrates this through inclusion of future features in the present-day study setting. Carefully staged studies concretise the features in high fidelity and make them realistic for the participants. This decreases the need for other staging techniques.

Propping is the most fundamental staging technique. It may be used in several ways. Propping helps to bring research settings a more authentic feel despite their artificial, futuristic characteristics. This helps the subjects suspend their awareness of being participants in a research study. In practice, every evaluative study involves propping, since the prototype is essentially a prop for a future technology. Other propping methods involve physical props (sometimes entire spaces, as in smart-house research), mockup content purposely created for the study, Wizard-of-Oz simulations of system functionality, and human assistants who enact the desired social settings without revealing to the participants that they are part of the research arrangements. In Figure 4, the resulting increase in authenticity is indicated with perfect instead of distorted circles also in the image's left pane.

In principle, propping can only improve the match between the study and the intended future. Propping is about creating a more likely research setting and more natural participant behaviour. The pitfalls of propping are related to the risk of propping the 'wrong' future and to the technique's costliness. It is expensive and could end up focused on irrelevant factors. This can restrict the number of participants or the duration of the evaluation. With very high-fidelity propping, the study may need additional control, because the participant may assume too much of the prototype's functionality and may try to carry out overly complicated actions.

Setting selection, *promotion*, and *repetition* are staging techniques that also involve an element of control. They make use of *exaggeration* and bring more future to the research setting than would be natural. This increases the likelihood of the participants encountering the features of interest sufficiently often during the study. Exaggeration is often necessary because the research questions at the heart of an evaluative study may be related to only a small part of the entire prototype or envisioned future. Therefore, researchers need to ensure that the study produces data about participants' interactions with these features especially. In Figure 4, the exaggerated feature is shown in an expanded form.

Setting selection involves exaggeration if the study is conducted in a setting wherein the behaviours of interest are more intensive or frequent for natural reasons. For example, Jacucci et al. [31] applied this in their mobile group-messaging study. They increased the messaging intensity by conducting the study at a mass sports event where frequent splitting into geographically separated sub-groups was inevitable. This increased the participants' needs for communication and, hence, use of the messaging prototype.

In feature promotion, in turn, participants are guided to pay attention to particular features of the prototype's UI through greater visual prominence of those features, them appearing more frequently, or emphasis on them in a tutorial on the technology before the empirical part of the study starts. The above-mentioned study of mobile messaging promoted use of participants' proximity information by showing it on every screen in the prototype's UI [31] and by emphasising this feature in the tutorial.

Finally, repetition causes participants to encounter the futuristic feature more often (see Figure 4). To achieve this effect, the researchers may use instruction or stage the setting in such a way that the desired repetition seems to occur naturally. For example, a group-communication prototype may include a researcher-generated within-app news feed that the participants might want to check frequently (cf. [31]). This encourages the participants to check their messages and thereby also to communicate with the tool more frequently.

In exaggeration, some features are accentuated relative to others. This technique therefore could result in fabricated observations. For example, if a user is tasked with posting messages repeatedly, the contents of these messages are not going to represent issues that a user would normally communicate about to others. From such a study, one cannot straightforwardly make claims as to what users wanted to communicate through the prototype. However, the study may be valid as an in-the-wild evaluation of a new text-entry method.

Exaggeration techniques' relationships with other control and staging techniques are complex. Additional staging may be needed if there are concerns about the exaggeration starting to become alienating. Less staging, in turn, may be needed if the selected setting keeps the participants captivated (as was seen

in the communication prototype study [31]). Use of a captivating setting may also result in less need for control. More control, on the other hand, may be needed if the participants could inadvertently exit the 'bubble' of the exaggerated research setting or if, despite feature promotion, they do not use the desired features frequently enough. The last approach was taken in the above-mentioned mobile AR map study through gamification elements added to the interface [43].

Recruiting is the last staging-oriented technique on our list. It is an essential part of all studies involving human participants, because a representative sample is fundamental to the validity of the study. In evaluative studies, recruiting can be used also for making the participants themselves more 'future-like' (see Figure 4). This is enabled by recruitment of lead users or experts as participants, thereby rendering the present-day setting closer to the future setting. For example, a camera manufacturer may recruit divers, downhill skiers, or skateboarders to participate in their evaluations because they are known as innovative communities in user-created video production. However, with lead users there may be a greater need for high-fidelity prototypes. Otherwise their skills cannot be fully capitalised upon. A highly skilled typist, for example, cannot be used to evaluate an input method if that method does not provide accurate tactile feedback. This imposes pressure on the staging, in terms of both the prototype implementation and the preparation of mockup material. Another issue is that experts, while representing the future users in one respect, may at the same time possess characteristics that do not generalise to the envisioned future user population. For example, in a comparison of navigation with 2D and 3D mobile maps [52], researchers recruited competitive players of the first-person-shooter game Counter-Strike to balance the comparison between well-mastered 2D maps and generally poorly mastered 3D maps. They considered Counter-Strike players likely to possess strong 3D navigation skills in addition to 2D ones. However, while the 2D and 3D navigation methods were thus rendered comparable, it cannot be ruled out that the participants' competitiveness and need for achievement were above average. This may have affected the findings, although such an effect could not be ascertained.

IMPLICATIONS FOR EVALUATORS

We started this paper by presenting the present–future gap: the tension between the features of an evaluative study, carried out in the present, and its actual use context in some yet-unknown future. We have now argued that evaluators take several decisions that concretise the intended future in the present-time study.

Our third major point listed in the introduction addresses the mindset that evaluators should adopt in their evaluative studies. Four questions for evaluators arise. Firstly, how should one design evaluative studies if methodological choices are speculative? Secondly, what does 'validity' mean when there is no ground truth (in the present) to consider? Thirdly, how can an evaluator think about the possible futures when designing studies? We conclude the paper with the fourth question: given all that has been said, how could studies be made more 'robust' against deviations from the intended future?

A New View on Study Design

Above, we have discussed how methodological choices place a study in various relationships with possible futures. The first implication we can highlight is that study design should be considered with this in mind. *Planning an evaluation involves several choices. From the future perspective, these are control- and staging-related choices, and each comes with unique considerations.*

Consider the above-mentioned study of 2D and 3D mobile maps [52]. Besides the recruiting already mentioned, it made use of narrowing, inhibition, repetition, and propping. The study was narrowed to address only navigation with a phone in situations wherein the map had to be used immediately. The participants were inhibited from experiencing some 'nuisance factors': uses of mobile phones that are commonplace in day-to-day life but would in this case have made it impossible to carry out the desired analyses. For example, switching between the phone's applications was not allowed during the study. Finally, the participants were also inhibited from wandering outside the area that had been 3D-modelled. The repetition in the study involved tasks that the participants had to carry out in a predetermined order. Finally, the prototype itself was a major propping effort. It used a semi-photorealistic model of a city centre that was optimised to run in an interactive mode on a smartphone [49]. With this study, the researchers were able to compare 2D and 3D navigation in a fair manner. However, this entailed diminished generalisability to non-hurried freeform urban exploration.

Techniques can and should, therefore, be used in combination. Table 1 lists some of the contingencies that come into play in this event; however, we have not yet looked at how an evaluator might make good choices in any given case.

Projective Validity: A New Validity Concept

The second implication is that what we mean by 'validity' must be rethought in the context of evaluative studies. Evaluative studies in HCI are speculative: they are empirical studies of something that 'might be'. Because of this, the validity of an evaluative study cannot be fully determined in the present. Hence, a serious question arises: are the traditional criteria [16, 23] for empirical research—internal, construct, conclusion, and external validity—sufficient for this purpose? If we consider them closely, none of the four types of validity directly captures the issue at stake:

Internal validity: Were the independent variables linked to the observed changes in the dependent variables? Internal validity, since it is related only to observations in a study, does not have future-orientation.

Construct validity: Is the independent variable the true cause of the change observed? In other words, to what degree does the test measure what it claims, or purports, to be measuring? This type of validity too is solely oriented to the present.

Conclusion validity: Was the observed change reliably measured? Normally, the conclusion in this respect is constrained to the present.

External validity: Does the observed change generalise to the target population, context, treatments, and outcomes? Here, the issue is that external validity should stretch into some future with its own populations and contexts.

In summary, only external validity is future-contingent. In search of a better name, we call the missing validity criterion *projective validity* and offer the following proposal: a study with high *projective validity* is 'future-proof' if the conclusions drawn hold in the relevant future contexts.

What the relevant future context is can vary greatly between studies. It may be very narrow or even implausible, depending on the researcher's intentions. This means that the projective validity of a study may be known only retrospectively, if ever. For example, studies of disaster-relief technologies may envision future settings that might never materialise. Also, a study may address desirable technologies that could be impossible to develop successfully (e.g., natural-speech input in noisy environments) or speculate on 'what-if's and undesirable futures (e.g., [5]). While evaluating projective validity would be important for HCI, doing so is often impossible. What is an HCI researcher to do, then? We offer the following solution.

Projective validity may be assessed, although subjectively, via justification of a '*margin of tolerance*' within which the study's projections should be considered valid. If the margin is large, the actual future (when it arrives) can deviate considerably from the study's concretised vision while the findings still hold water. Alternatively, if the margin of tolerance is narrow, the study will be 'brittle' in the face of the range of possible futures, since there is only a small likelihood that the actual future will be exactly as envisioned in the study. In this case, the value of the study comes from suggestions as to the direction that the future of HCI should take, either towards or away from the study's envisioned future.

There is an exception to the above: an evaluative study may, in fact, 'create' a future—if the findings are insightful enough. Knowledge about a desirable future may change our perception of it and make it more likely to occur via some mechanisms beyond the evaluator's control. For example, Weiser's [70] and Ishii's [30] visionary explorations of interactions beyond the desktop in the 1990s dramatically shaped research trends in HCI. By basing the study's projections on a well-considered margin of tolerance, researchers are better equipped to argue for the importance of their findings.

Defining Intended Futures

Our third implication is that all evaluations should be planned with explicated futures in mind. The concept of a margin of tolerance is useful here. The wider the margin of plausible futures that the empirical evaluation targets, the better. Alternatively, one can think about it in terms of expected probability that the evaluation matches with the futures concretised in the study. Defining a tolerance margin therefore consists of defining 1) a timeframe (duration until future), 2) a list of those features that it is sufficient and necessary to stage or control, 3) the details of each such feature, and 4) a feature list with specifications for invariant present-day features that are not assumed to change during the timeframe.

Let us consider an example: an evaluation of a digital note-taking system, assuming a five-year timeframe and targeting a future with tablet-like devices with a stylus that have a minimum of 150 DPI resolution and a 5 ms tracking delay. The device should reach 97% accurate recognition for hand-written text after initial training. It would be part of a digital ecosystem wherein note-sharing would be a seamless activity. Given this, the present-day features to be controlled (via, for example, removal) would involve, among others, scrap paper and ordinary pens. The invariant features would include, for example, the limitations imposed by high energy consumption of hand-held devices. Let us assume that one finding from such a study would be that users start sharing more and more third-party material with their own annotations. This could be projected to imply requirements for the ecosystem and its digital rights management (DRM) policies. With proper specification of the ecosystem at the outset of the study, the implications would be sharp, pointing out clear differences from the present-day ecosystems and DRM policies. The margin would thus offer a sharper interpretive lens on the empirical findings.

Research in the policy and strategy arenas has already suggested something akin to the tolerance margin. These disciplines develop forecasts differently in accordance with the level of uncertainty. Figure 5, adapted from this literature [69], depicts a typology with five levels, which range from almost complete certainty to two forms of 'deep uncertainty'—futures whose prediction models, probabilities, and relative desirability levels experts cannot even agree about [38]. In this end, the distance of one future from another is unknown, so neighbouring futures cannot be postulated. At the other end of the continuum (at the left in Figure 5), in turn, the uncertainties are low and are usually estimated numerically. The corresponding margin of tolerance is then a parameter range for only one type of future. Figure 5 illustrates this with green stripes. This corresponds to near-future evaluations. An evaluation of a distant future, in contrast, involves deep uncertainty wherein the margin of tolerance may disappear altogether: there are only unique futures, each one impossible to compare to another. Here, the researcher must choose the future whose evaluation seems most informative in itself. The findings obtained from such a study are not generalisable to the other futures but can prove 'existence' of at least one possible future wherein a given prototype has a meaningful role in human life.

The required precision of the margin therefore depends on the uncertainty of the future. A failure to specify margins would leave many of the staged or controlled features vague or unspecified. This would leave much space for wishful thinking in which the researchers interpret the empirical results selectively, in the worst case disregarding inconvenient observations. Many HCI studies focus on the middle range along this spectrum, with a goal of studying one among many alternative futures. In policy and strategy research, scenario-based techniques are common at that level [27, 60]. The forecasts are based on trends that have been cross-validated across multiple sources, to guard against overconfidence [60]. Instead of using trends only as an inspiration [59], taking an approach

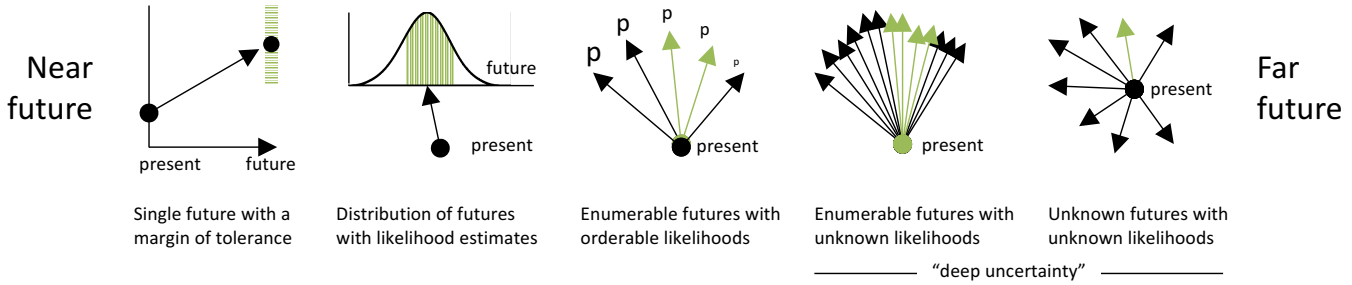

Figure 5. Levels of uncertainty involved with future scenarios. We indicate the margin of tolerance in green.

similar to this would improve the validity of HCI's evaluative studies.

Near the deep uncertainty end of the spectrum, design fiction (e.g., [39]) comes closest to use of a margin of tolerance, with its fictitious stories about possible future HCI. This method, however, has been used mostly in envisioning, not in projection. If design fiction were used in projection and forced to stay within a given margin of tolerance, its imaginative strength would not be exploited to its fullest extent. Whether this hindrance is a problem remains an open question.

CONCLUSION: TOWARDS A NEW MINDSET

We have drawn attention to a pervasive but overlooked characteristic of evaluative HCI studies: their future-oriented nature. An evaluation of a prototype is a study of 'what might be'. This is a potentially significant problem, because it may hinder HCI studies' ability to inform of how interactive technology should be designed. The traditional 'go-to disciplines' that have shaped our evaluation methodology—the behavioural and social sciences in particular—are ill-equipped for providing methodological guidance on evaluations that involve future-orientation. Therefore, if we are to improve the quality of our usability studies, field trials, and lab evaluations, the HCI community needs to start developing its own approach. This paper constitutes an attempt to increase awareness of this unique methodological challenge.

Before turning to guidance for evaluative praxis, a reminder of the limitations of our argument is in place. We have focused on analysing the implications of future-orientedness only within that framework for HCI research that emphasises validity. Similar critical analysis of the future-orientedness of HCI's third paradigm would be a desirable complement to our work. We have exposed the complexity of future-contingency and offered mostly conceptual tools for reflection. Moreover, our contribution is most applicable to evaluations of standalone systems. This limitation may be more acceptable in a research context but could be a problem in a company context, especially in the evaluation of business infrastructure.

When deliberations about possible futures and their effects on methodological choices become more commonplace in HCI research, our field will collectively improve the projective validity of our evaluations. For our fourth and final take-away point, summarising the discussion above, we enumerate five practices that could help practitioners improve in this regard.

1. Mindset: Evaluators should gear their studies such that the primary goal is to understand the prototype in plausible future states of affairs instead of the present.

2. Reflection: Evaluators should consider the methodological options in terms of staging (adding future-like features) and control (inhibiting features that conflict with the future). The critical evaluation of methodological choices ranges from the specifics of the prototype's construction through participant recruitment to study design, all from the perspective of projective validity.

3. Replication: The purpose of replication studies should be rethought. Replication in HCI is normally thought of in terms of increasing the confidence associated with a result. Through replication *with the future in mind*, one can make a result more 'future-proof', by proving that said finding holds over a larger margin of tolerance than was anticipated in the original study. Alternatively, replication may test boundary conditions of previous studies whose margins of tolerance seem broader than can be validly justified.

4. Transparency: Evaluators should try to make their assumptions about the future explicit. This includes explicating assumptions about the targeted futures and identifying uncertainties, as well as discussing how the study at hand concretises the intended futures.

5. Post-launch monitoring: While many companies (for example, medical companies) routinely engage in assessing how well the products they have launched fare in the market, such analyses are rarely carried out with the goal of improving the ability to design better evaluative studies. Learning from past studies and their projective validity, once that validity can be ascertained, would improve the evaluative studies and their projections.

We believe that adopting these practices would help researchers structure their studies and pay attention to elements that increase the validity and insightfulness of those studies.

ACKNOWLEDGEMENTS

AS has received funding from the Academy of Finland (grants 259281 and 298879). AO has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant 637991). We thank David McGookin, Barry Brown, Duncan Brumby, Kasper Hornbæk, Mika Jokiniemi, Stuart Reeves, Bariş Serim, and John Zimmerman for comments.

REFERENCES

1. Gregory D. Abowd. 1999. Classroom 2000: An Experiment with the Instrumentation of a Living Educational Environment. *IBM Systems Journal* 38, 4 (1999), 508–530. DOI: http://dx.doi.org/10.1147/sj.384.0508
2. Evan Autherton. 2016. Science Fiction Prototyping at Work. *IEEE Computer* 49, 8 (2016), 109–111. DOI: http://dx.doi.org/10.1109/MC.2016.229
3. Liam Bannon. 1991. From Human Factors to Human Actors: The Role of Psychology and Human–Computer Interaction Studies in System Design. In *Design at Work: Cooperative Design of Computer Systems*, Joan M. Greenbaum and Morten Kyng (Eds.). Lawrence Erlbaum, Hillsdale, NJ, Chapter 2, 25–44.
4. Isabelle Baszanger and Nicolas Dodier. 2004. Ethnography: Relating the Part to the Whole. In *Qualitative Research: Theory, Methods and Practice* (2 ed.), David Silverman (Ed.). Sage Publications, London, UK, Chapter 2, 9–34.
5. Steve Benford, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, and Tom Rodden. 2012. Uncomfortable Interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012)*. ACM Press, New York, NY, 2005–2014. DOI: http://dx.doi.org/10.1145/2207676.2208347
6. J. Bergström-Lehtovirta, A. Oulasvirta, and S. Brewster. 2011. The Effects of Walking Speed on Target Acquisition on a Touchscreen Interface. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI 2011)*. ACM Press, New York, NY, 143–146. DOI: http://dx.doi.org/10.1145/2037373.2037396
7. Wiebe E. Bijker, Thomas P. Hughes, and Trevor J. Pinch (Eds.). 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. The MIT Press, Cambridge, MA.
8. Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI Interprets the Probes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007)*, Mary Beth Rosson and David Gilmore (Eds.). ACM Press, New York, NY, 1077–1086. DOI: http://dx.doi.org/10.1145/1240624.1240789
9. Laurens Boer and Jared Donovan. 2012. Provotypes for Participatory Innovation. In *Proceedings of the Designing Interactive Systems Conference (DIS 2012)*. ACM, New York, NY, 388–397. DOI: http://dx.doi.org/10.1145/2317956.2318014
10. Robert O. Briggs, Gert-Jan De Vreede, and Jay F. Nunamaker, Jr. 2003. Collaboration Engineering with ThinkLets to Pursue Sustained Success with Group Support Systems. *Journal of Management Information Systems* 19, 4 (2003), 31–64.
11. Barry Brown. 2002. Studying the Use of Mobile Technology. In *Wireless World: Social and Interactional Aspects of the Mobile Age*, Barry Brown, Nicola Green, and Richard Harper (Eds.). Springer, London, UK, Chapter 1, 3–15. DOI: http://dx.doi.org/10.1007/978-1-4471-0665-4_1
12. Barry Brown, Stuart Reeves, and Scott Sherwood. 2011. Into the Wild: Challenges and Opportunities for Field Trial Methods. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011)*. ACM, New York, NY, 1657–1666. DOI: http://dx.doi.org/10.1145/1978942.1979185
13. Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI 2016)*. ACM Press, New York, NY, 981–992. DOI: http://dx.doi.org/10.1145/2858036.2858498
14. John M. Carroll. 2000. *Making Use: Scenario-Based Design of Human–Computer Interactions*. The MIT Press, Cambridge, MA.
15. John. M. Carroll and Wendy. A. Kellogg. 1989. Artifact As Theory-Nexus: Hermeneutics Meets Theory-Based Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1989)*, K. Bice and C. Lewis (Eds.). ACM Press, New York, NY, 7–14. DOI: http://dx.doi.org/10.1145/67449.67452
16. Thomas D. Cook and Donald Thomas Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin, New York, NY.
17. Ron Eglash, Jennifer L. Croissant, Giovanna Di Chiro, and Rayvon Fouché (Eds.). 2004. *Appropriating Technology: Vernacular Science and Social Power*. University of Minnesota Press, Minneapolis, MN.
18. Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Cultural Probes. *Interactions* 6, 1 (1999), 21–29. http://doi.org/10.1145/291224.291235
19. William Gaver. 2012. What Should We Expect from Research through Design?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012)*. ACM Press, New York, NY, 937–946. DOI: http://dx.doi.org/10.1145/2207676.2208538
20. Steve Harrison, Phoebe Sengers, and Deborah Tatar. 2011. Making Epistemological Trouble: Third-Paradigm HCI As Successor Science. *Interacting with Computers* 23, 5 (2011), 385–392. DOI: http://dx.doi.org/10.1016/j.intcom.2011.03.005
21. Steve Harrison, Deborah Tatar, and Phoebe Sengers. 2007. The Three Paradigms of HCI. In *Proceedings of alt.chi 2007*.

22. Thomas T. Hewett, Ronald Baecker, Stuart Card, Tom Carey, Jean Gasen, Marilyn Mantei, Gary Perlman, Gary Strong, and William Verplank. 1992. *ACM SIGCHI Curricula for Human–Computer Interaction*. Technical Report. New York, NY.
23. Kasper Hornbæk. 2013. Some Whys and Hows of Experiments in Human–Computer Interaction. *Foundations and Trends in Human–Computer Interaction* 5, 4 (2013), 299–373. DOI: http://dx.doi.org/10.1561/1100000043
24. Steven Houben, Jakob E. Bardram, Jo Vermeulen, Kris Luyten, and Karin Coninx. 2013. Activity-Centric Support for Ad Hoc Knowledge Work—a Case Study of Co-Activity Manager. In *Proceedings of the SIGCHI Conference on Human Factors in Computing (CHI 2013)*. ACM Press, New York, NY, 2263–2272. DOI: http://dx.doi.org/10.1145/2470654.2481312
25. Edwin Hutchins. 1995. *Cognition in the Wild*. The MIT Press, Cambridge, MA.
26. Hilary Hutchinson, Wendy Mackay, Bosse Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, Björn Eiderbäck, Sinna Lindquist, and Yngve Sundblad. 2003. Technology Probes: Inspiring Design for and with Families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2003)*. ACM Press, New York, NY, 17–24. DOI: http://dx.doi.org/10.1145/642611.642616
27. J. Hyndman and George Athanasopoulos. 2013. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia.
28. Giulio Iacucci and Kari Kuutti. 2002. Everyday Life As a Stage in Creating and Performing Scenarios for Wireless Devices. *Personal and Ubiquitous Computing* 6, 4 (2002), 299–306. DOI: http://dx.doi.org/10.1007/s007790200031
29. Giulio Iacucci, Kari Kuutti, and Mervi Ranta. 2000. On the Move with the Magic Thing: Role Playing in Concept Design of Mobile Services and Devices. In *Proceedings of the Conference on Designing Interactive Systems (DIS 2000)*. ACM Press, New York, NY, 193–202. http://doi.org/10.1145/347642.347715
30. Hiroshi Ishii and Brygg Ullmer. 1997. Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 1997)*. ACM Press, New York, NY, 234–241. DOI: http://dx.doi.org/10.1145/258549.258715
31. Giulio Jacucci, Antti Oulasvirta, Tommi Ilmonen, John Evans, and Antti Salovaara. 2007. CoMedia: Mobile Group Media for Active Spectatorship. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007)*, Mary Beth Rosson and David Gilmore (Eds.). ACM Press, New York, NY, 1273–1282. DOI: http://dx.doi.org/10.1145/1240624.1240817
32. Jesper Kjeldskov and Jeni Paay. 2012. A Longitudinal Review of Mobile HCI Research Methods. In *Proceedings of the 14th International Conference on Human–Computer Interaction with Mobile Devices and Services (MobileHCI 2012)*. ACM Press, New York, NY, 69–78. DOI:http://dx.doi.org/10.1145/2371574.2371586
33. Jesper Kjeldskov and Mikael B. Skov. 2014. Was It Worth the Hassle? Ten Years of Mobile HCI Research Discussions on Lab and Field Evaluations. In *Proceedings of the 16th International Conference on Human–Computer Interaction with Mobile Devices & Services (MobileHCI 2014)*, Aaron Quigley, Sara Diamond, Pourang Irani, and Sriram Subramanian (Eds.). ACM Press, New York, NY, 43–52. DOI: http://dx.doi.org/10.1145/2628363.2628398
34. Jesper Kjeldskov, Mikael B. Skov, Benedikte S. Als, and Rune T. Høegh. 2004. Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Proceedings of the 6th International Symposium on Mobile Human Computer Interaction (MobileHCI 2004) (Lecture Notes in Computer Science)*, Stephen A. Brewster and Mark D. Dunlop (Eds.), Vol. LNCS 3160. Springer, Berlin and Heidelberg, Germany, 61–73. DOI: http://dx.doi.org/10.1007/978-3-540-28637-0_6
35. Sari Kujala and Marjo Kauppinen. 2004. Identifying and Selecting Users for User-Centered Design. In *Proceedings of the Third Nordic Conference on Human–Computer Interaction (NordiCHI 2004)*, Roope Raisamo (Ed.). ACM Press, New York, NY, 297–303. DOI:http://dx.doi.org/10.1145/1028014.1028060
36. Kari Kuutti and Liam J. Bannon. 2014. The Turn to Practice in HCI: Towards a Research Agenda. In *Proceedings of the SIGCHI Conference on Human Factors in Computing (CHI 2014)*, Matt Jones, Philippe Palanque, Albrecht Schmidt, and Tovi Grossman (Eds.). ACM Press, New York, NY, 3543–3552. DOI: http://dx.doi.org/10.1145/2556288.2557111
37. Morten Kyng and Joan Greenbaum (Eds.). 1991. *Design at Work: Cooperative Design of Computer Systems*. Lawrence Erlbaum, Hillsdale, NJ.
38. Robert J. Lempert, Steven W. Popper, and Steven C. Bankes. 2003. *Shaping the Next Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*. Technical Report MR-1626. The RAND Pardee Center, Santa Monica, CA. http://www.rand.org/pubs/monograph_reports/MR1626.html
39. Joseph Lindley and Paul Coulton. 2016. Pushing the Limits of Design Fiction: The Case for Fictional Research Papers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI 2016)*, Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade (Eds.). ACM Press, New York, NY, 4032–4043. DOI: http://dx.doi.org/10.1145/2858036.2858446

40. Daniel MacKenzie and Judy Wajcman (Eds.). 1998. *The Social Shaping of Technology* (2 ed.). Open University Press, Buckingham, UK.

41. Katri Mehto, Vesa Kantola, Sauli Tiitta, and Tomi Kankainen. 2006. Interacting with User Data—Theory and Examples of Drama and Dramaturgy As Methods of Exploration and Evaluation in User-Centered Design. *Interacting with Computers* 18, 5 (2006), 977–995. DOI:http://dx.doi.org/10.1016/j.intcom.2006.05.006

42. David R. Millen. 2000. Rapid Ethnography: Time Deepening Strategies for HCI Field Research. In *Proceedings of the Conference on Designing Interactive Systems (DIS 2000)*. ACM Press, New York, NY, 280–286. DOI:http://dx.doi.org/10.1145/347642.347763

43. Ann Morrison, Antti Oulasvirta, Peter Peltonen, Saija Lemmelä, Giulio Jacucci, Gerhard Reitmayr, Jaana Näsänen, and Antti Juustila. 2009. Like Bees around the Hive: A Comparative Study of a Mobile Augmented Reality Map. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009)*. ACM Press, New York, NY, 1889–1898. DOI: http://dx.doi.org/10.1145/1518701.1518991

44. Bonnie A. Nardi. 1993. *A Small Matter of Programming: Perspectives on End User Computing*. The MIT Press, Cambridge, MA.

45. Gina Neff and David Stark. 2004. Permanently Beta: Responsive Organization in the Internet Era. In *Society Online: The Internet in Context*, Philip N. Howard and Steve Jones (Eds.). Sage Publications, Thousand Oaks, CA, Chapter 11, 173–188. DOI: http://dx.doi.org/10.4135/9781452229560.n11

46. Ulric Neisser. 1976. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. W. H. Freeman and Company, San Francisco, CA.

47. Jakob Nielsen. 1993. *Usability Engineering*. Academic Press, Boston, MA.

48. Jay F. Nunamaker, Jr., Robert O. Briggs, Daniel D. Mittleman, Douglas R. Vogel, and Pierre A. Balthazard. 1996. Lessons from a Dozen Years of Group Support Systems Research: A Discussion of Lab and Field Findings. *Journal of Management Information Systems* 13, 3 (1996), 163–207.

49. Antti Nurminen. 2008. Mobile 3D City Maps. *IEEE Computer Graphics and Applications* 28, 4 (2008), 20–31. DOI:http://dx.doi.org/10.1109/MCG.2008.75

50. William Odom, John Zimmerman, Scott Davidoff, Jodi Forlizzi, Anind Dey, and Min-Kyung Lee. 2012. A Fieldwork of the Future with User Enactments. In *Proceedings of the Designing Interactive Systems Conference (DIS 2012)*. ACM Press, New York, NY, 338–347. DOI: http://dx.doi.org/10.1145/2317956.2318008

51. Nelly Oudshoorn and Trevor Pinch (Eds.). 2003. *How Users Matter: The Co-Construction of Users and Technology*. The MIT Press, Cambridge, MA.

52. Antti Oulasvirta, Sara Estlander, and Antti Nurminen. 2009. Embodied Interaction with a 3D Versus 2D Mobile Map. *Personal and Ubiquitous Computing* 13, 4 (2009), 303–320. DOI: http://dx.doi.org/10.1007/s00779-008-0209-0

53. Antti Oulasvirta and Kasper Hornbæk. 2016. HCI Research As Problem-Solving. In *Proceedings of the SIGCHI Conference on Human Factors in Computing (CHI 2016)*, Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade (Eds.). ACM Press, New York, NY, 4956–4967. DOI: http://dx.doi.org/10.1145/2858036.2858283

54. Antti Oulasvirta, Esko Kurvinen, and Tomi Kankainen. 2003. Understanding Contexts by Being There: Case Studies in Bodystorming. *Personal and Ubiquitous Computing* 7, 2 (2003), 125–134. DOI: http://dx.doi.org/10.1007/s00779-003-0238-7

55. Stuart Reeves. 2012. Envisioning Ubiquitous Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing (CHI 2012)*. ACM Press, New York, NY, 1573–1582. DOI: http://dx.doi.org/10.1145/2207676.2208278

56. Stuart Reeves, Murray Goulden, and Robert Dingwall. 2016. The Future as a Design Problem. *Design Issues* 32, 3 (2016). DOI: http://dx.doi.org/10.1162/DESI_a_00395

57. Yvonne Rogers, Kay Connelly, Lenore Tedesco, William Hazlewood, Andrew Kurtz, Robert E. Hall, Josh Hursey, and Tammy Toscos. 2007. Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. In *Proceedings of the 9th International Conference on Ubiquitous Computing (Ubicomp 2007) (Lecture Notes in Computer Science)*, John Krumm, Gregory D. Abowd, Aruna Seneviratne, and Thomas Strang (Eds.), Vol. LNCS 4717. Springer, Berlin Heidelberg, 336–353. DOI: http://dx.doi.org/10.1007/978-3-540-74853-3_20

58. Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests* (2 ed.). Wiley Publishing, Indianapolis, IN.

59. Antti Salovaara and Petri Mannonen. 2005. Use of Future-Oriented Information in User-Centered Product Concept Ideation. In *Proceedings of the IFIP TC13 International Conference on Human–Computer Interaction (Interact 2005) (Lecture Notes in Computer Science 3585 (LNCS))*. Springer-Verlag, Berlin, Germany, 727–740. DOI: http://dx.doi.org/10.1007/11555261_58

60. Paul J. H. Schoemaker. 1991. When and How to Use Scenario Planning: A Heuristic Approach with Illustration. *Journal of Forecasting* 10, 6 (1991), 549–564. DOI: http://dx.doi.org/10.1002/for.3980100602

61. Donald A. Schön. 1983. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, New York, NY.
62. Douglas Schuler and Aki Namioka (Eds.). 1993. *Participatory Design: Principles and Practices*. Lawrence Erlbaum, Hillsdale, NJ.
63. W. Shadish, T. Cook, and D. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, MA.
64. Herbert A. Simon. 1969. *The Sciences of the Artificial*. The MIT Press, Cambridge, MA.
65. Erik Stolterman and Mikael Wiberg. 2010. Concept-Driven Interaction Design Research. *Human–Computer Interaction* 25, 2 (2010), 95–118.
66. Lucy A. Suchman. 1987. *Plans and Situated Actions: The Problem of Human–Machine Communication*. Cambridge University Press, Cambridge, UK.
67. Eric von Hippel. 1988. *The Sources of Innovation*. Oxford University Press, New York, NY.
68. Eric von Hippel. 2005. *Democratizing Innovation*. The MIT Press, Cambridge, MA.
69. Warren E. Walker, Robert J. Lempert, and Jan H. Kwakkel. 2013. Deep Uncertainty. In *Encyclopedia of Operations Research and Management Science*, Saul I. Gass and Michael C. Fu (Eds.). Springer, New York, NY, 395–402. DOI: http://dx.doi.org/10.1007/978-1-4419-1153-7_1140
70. Mark Weiser. 1991. The Computer for the 21st Century. *Scientific American* 265, 3 (1991), 66–73.
71. William Foote Whyte (Ed.). 1990. *Participatory Action Research*. Sage Publications, Thousand Oaks, CA.
72. Terry Winograd, John Bennett, Laura De Young, and Bradley Hartfield (Eds.). 1996. *Bringing Design to Software*. ACM Press, New York, NY.
73. Terry Winograd and Fernando Flores. 1986. *Understanding Computers and Cognition: A New Foundation for Design*. Addison-Wesley, Reading, MA.
74. Salu Ylirisku, Virtu Halttunen, Johanna Nuojua, and Antti Juustila. 2009. Framing Design in the Third Paradigm. In *Proceedings of the SIGCHI Conference on Human Factors in Computing (CHI 2009)*, Dan R. Olsen, Jr, Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott Hudson, and Saul Greenberg (Eds.). ACM Press, New York, NY, 1131–1140. DOI: http://dx.doi.org/10.1145/1518701.1518874
75. John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007a. Research through Design As a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007)*. ACM Press, New York, NY, 493–502. DOI: http://dx.doi.org/10.1145/1240624.1240704
76. John Zimmerman, Erik Stolterman, and Jodi Forlizzi. 2007b. An Analysis and Critique of Research through Design: Towards a Formalization of a Research Approach. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS 2007)*. ACM Press, New York, NY, 310–319. DOI: http://dx.doi.org/10.1145/1858171.1858228