Into the Wild: Challenges and Opportunities for Field Trial Methods

Barry Brown[1 2], **Stuart Reeves**[3] **and Scott Sherwood**[4]

[1] Department of Communication University of California San Diego barry@ucsd.edu

[2] Mobile Life VINN Excellence center Stockholm Sweden

[3] Horizon Digital Economy Research University of Nottingham, UK stuart@tropic.org.uk

[4] Department of Computing Science University of Glasgow UK sherwood@dcs.gla.ac.uk

ABSTRACT

Field trials of experimental systems 'in the wild' have developed into a standard method within HCI - testing new systems with groups of users in relatively unconstrained settings outside of the laboratory. In this paper we discuss methodological challenges in running user trials. Using a 'trial of trials' we examined the practices of investigators and participants - documenting 'demand characteristics', where users adjust their behaviour to fit the expectations of those running the trial, the interdependence of how trials are run and the result they produce, and how trial results can be dependent on the insights of a subset of trial participants. We develop three strategies that researchers can use to leverage these challenges to run better trials.

Author Keywords

Field trials, methods, demand characteristics, ethnography.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Design, Experimentation, Human Factors.

INTRODUCTION

Within HCI perhaps one of the most popular methods for exploring novel technologies has been the system field trial or user study, where new systems are deployed and studied in relatively uncontrolled settings. From early on in HCI, researchers have built systems and given them to users to test in various different ways [3, 16, 28]. While early tests were usually in controlled laboratory settings, the field trial located 'in the wild' has emerged to address both diverse settings of use as well as interest in the 'unanticipated use' of systems [25]. The user trial, deployment or study has characteristics that distinguish it from other forms of deployment. While trials often involve an experimental design, there is usually a bias towards unstructured, more naturalistic experiments. Results span a mix of the qualitative and quantitative, with a open orientation towards finding out 'what happens' and drawing design principles or recommendations about users' reactions.

This paper seeks to draw lessons for how we might better conduct and present user trials in HCI. We document one trial of a photo sharing application - but not as an investigation into a system that needed to be tested but rather as a way of reflecting on user behaviour in system trials. Drawing on this trial we document three key issues that influence what happens in user trials yet are mostly absent from discussions of trial methods in HCI. *Demand characteristics* are where users in trials adjust and report on their behaviour in ways that fit with their perception of investigators' expectations. As an example, in our trial, participants discussed how they endeavoured to use our system so as to be able to be able to 'give' to those running the trial 'the results you are looking for'. *Lead participants* are where there is a reliance in reporting results from a trial on a small atypical subset of users who engage with and offer particular insight into the behaviour under investigation. In the trial discussed here one user took it on himself to organise and encourage use by others in the trial as well as offering to us insightful comments about the use of the system. Lastly *interdependence of methods and results*, describes how the way in which a particular trial is run, and the questions asked by investigators, intimately interact to narrow results and behaviour. In particular, the relationships between experimenter and participant can play as important a role in what happens in a trial as the design of the system.

We use this trial to explore how behaviours observed in trials are a product of the methods used, the orientation of investigators, interactions between participants, as well as the particular design of the system deployed. While this will come as no surprise to experienced researchers, foregrounding them allows us to explore how improve trial methodology - how we write up trials, but also how methods might better engage with the realities of trials as

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

they are run. Our goal is to not see these issues as 'problems' with trials as such, but rather as unavoidable characteristics of trials that take place outside controlled settings, and as characteristics that can be leveraged to improve trial practice.

What are field trials?

Field trials have become a common method for studying the use of novel technologies, widely used in fields (such as HCI, Ubicomp and CSCW) where the interests of investigators goes beyond technical feasibility to exploring user understandings, practices and the eventual uses that systems might be put. What we refer to as *trials* in this paper go under a range of different names - field experiments, deployments, evaluations, field studies, technical probes - but they share a set of common features. A new system, usually developed by the researchers, is given to a set of users who are asked (often implicitly) to use the system 'naturally' outside the laboratory the system was designed in. They then use the system for anything from a few hours to a year, frequently as part of their day to day life, or with the system deployed in their home or workplace. In more experimental trials users' behaviour may be constrained with tasks set for them to carry out, while other trials eschew such controls and attempt to encourage 'natural' use. In some cases trials are run as an extension of experiments, whereas for others a more ethnographic analytic mode is adopted.

All trials involve data collected from system use, be that in the form of system logs, user interviews, observations or user reports (e.g., experience sampling). Papers focus on the technology's use, with an implicit understanding that 'the results' are produced by the users and the technology, and are not simply a function of the trial. Some go further and make claims that, based on what happens in the trial, findings can be extended to understand how similar technologies might be used if they were released commercially, or to explicitly seek to evaluate the prototype positively or negatively. There is a common interest in studying user behaviour with a prototype system so as to understand how to better design later technologies. Three diverse examples of user trials are Lee & Day's Lifelogging trial [21], Bell et al.'s Feeding Yoshi [3], or Pousman et al.'s Tableau Machine [25].

FIELD TRIALS IN THE LITERATURE

Some of the earliest technology trials come from the first half of the 1990s, as research interests migrated from testing technology in laboratory experiments to more broadly investigating interactions with technology [2]. Such was the fragile nature of first prototypes that the earliest examples are actually investigations based in research labs: Portholes at EuroPARC [11], the Active Badge system [16], and in Bell Labs various mediaspace systems [28]. Even though researchers deployed the systems within their own workplaces, their reflections offered much in terms of forming an understanding of what happened when these technologies became more widespread.

Following these pioneering studies, trials and naturalistic deployments of systems have become a core method for investigating user interactions with systems (e.g. [4, 29]). In particular, with ubiquitous computing systems, the close coupling of environment, and the concern for systems that engage with users' everyday lives, has caused the field trial to flourish as a 'standard method' (e.g. [3, 5, 27]). Indeed, the nature of ubicomp limits the power of classical laboratory studies - for example, the accuracy of location tracking can be tested in relatively controlled setting, but a study of this type tells us little about user acceptance, or how tracking might work with an individual's routine mobility. Mobility and the embedding of systems into the environment encourages deployment outside laboratories, and interest in the interweaving of technology with everyday life encourages longer trials. Of late, with the lowering of the technological hurdles for testing and deploying new applications and hardware, field trials of technology 'in the wild' lasting weeks or months have become increasingly commonplace. Within HCI the home in particular has been one important site for field studies. Situating technology within the home and exploiting the unconstrained and unanticipated use that is to be found there, researchers have sought to use them as seeding technologies and inspirations for design [26]. Taking stock of the inherent difficulties posed by such deployments, aspects of the home have also been further modified and brought back to the laboratory. The rise of 'living labs' such as MITs PlaceLab [18] enabled researchers to develop spaces in which uncertainty found in home settings could be more managed, and yet retain many aspects of field trials.

While trials have been a popular method, discussion of the field trial as a method has been relatively absent, certainly when compared to debates around other methods for examining user interactions, such as ethnography [9]. Debates around the nature of 'probes' of certain forms, particularly Hutchinson et al.'s technology probes [17] and Gaver and Rabe's cultural probes [13], are perhaps the main exceptions. Both of these approaches share a notion of a lightweight technological intervention that attempts to investigate current practice and experience to inform the design of new artefacts. Technology probes introduce new technologies that support new practices, whereas cultural probes are "a design orientated way to acquire inspirational glimpses of communities targeted for design" [6].

One radical view of user trials is that such deployments act as a kind of "breaching experiment" [9]. Crabtree argues that user trials need grapple with "the absence of practice", meaning that researchers are attempting to support through technology a set of practices which have not yet emerged, leaving us questioning what within a trial we are actually studying. Crabtree describes technology trials as "breaching experiments" - an attempt to disturb existing,

taken-for-granted practices of doing things, so as to reveal how those things are done. To us this approach understates the power of trials; technology does not only disrupt but also potentially support new practices. We find in Harper et al.'s discussion of Active Badges [18] or Dourish and Adler's study of Portholes [12] more than simply the disruption of existing practice. Relatedly, Tolmie and Crabtree discuss the role of field trials in home environments [30]. Through attending to the deployment of a novel home-based technology (a video feed from a camera on a pole outside the home), they detailed how the trial participants and researchers oriented toward the technology and one another. They argue that the introduction of technology into the home disrupts 'everyday' domestic life, limiting the potential for understanding the 'everyday', something they see as the target of their studies. In contrast, Carter's et al.'s paper 'Exiting the cleanroom' [7] focuses on how we might better run trials so as to make them more 'ecologically valid' and to explore the barriers to more effective technology trials as a mode of design.

Finally, some discussions of field trials have challenged the utility of the practice itself and what it contributes to research. Davies [10] asserts that field trial-style research is costly in terms of time, but also is frequently the wrong tool for the job. If the goal is to prove that something technical is possible then usually there is no experiential concept to actually be proven - the demonstrator proves something that is already widely known, or that this point would be better demonstrated theoretically or experimentally. While Davies' points are more relevant to the specific technical goals of user trials, Kjeldskov argues that field trials provide little added user studies value, again emphasising that they are labour-intensive and that much that is found out would be better discovered in a laboratory study [20]. Whittaker et al [31] go on to argue that we should focus on constrained reference tasks and that while field trials have a role they should be used more to "modify existing task definitions for future evaluations". Rogers presents a robust response to these points demonstrating how many usability problems do not arise in the laboratory when compared to in-situ use, but also how field trials "provide a contextual backdrop against which to reflect upon the design of user experience and the mobile device sensitising us to how [our system] would (rather than should) be used in practice" [27].

EXAMINING TRIALS WITH A TRIAL

This variety of approaches towards system field trials demonstrates some of the diverse understandings of what field trials can do. While these discussions engage in productive ways with the utility of trials, or what claims can be made, our interest was in exploring what is *not* traditionally reported in trails papers - the messy details of trial practice that seldom go reported. Accordingly, we developed a trial of sorts to explore how the 'non-designed' features of a trial affected proceedings. Our goal was not to evaluate or critique trials as a method, but rather to use a trial to illustrate what we would argue are unavoidable characteristics of real world technology trials, so as to throw into relief our experiences from previous trials (e.g., [1, 3]).

Our trial had many of the trappings of a conventional study - we interviewed participants, observed them using the system and collected log data of the system's use. Just as we might use statistics to evaluate a statistical study, or ethnography to study ethnography, our goal was to use a trial to study what influences behaviour in user trials. Our interview questions, analysis and writeup focused on questions of what motivated participants to behave certain ways, and how the interactions between participants and ourselves influenced behaviour. The focus in the design and analysis was on how different interactions between participants, and between ourselves as investigators, had a role to play in participants' behaviour.

We tested an iPhone-based photo sharing system,that allowed users to distribute photos amongst a group alongside commenting on each photo. In some senses this application was fairly prosaic. The participants came from a variety of different professions and backgrounds. 20 participants were involved in our trial, split into four separate groups. As the composition of the groups, and the relationships between those involved were of particular interest for our study, we will describe these in some detail (Table 1).

Firstly, our groups varied in terms age range, and of whether they self-identified as football fans. The table also indicates the differing recruitment methods. For Groups A and D, recruitment was performed through a 'friends of friends' method (A was recruited through a family member of one of the authors, D via friends of one of the authors). Group B was recruited through University clubs, via membership of one of the authors, and Group C was recruited through press coverage, resulting in one of the members of the group contacting the authors to become a participant (and in turn recruiting his friends). Naturally, each group had a particular social configuration, which was of relevance to the way in which interaction played out during the trial. **Group A** were socially close, often exchanged SMS messages, and regularly engaged in communication together via social networking sites. For Group A we also had a 'primary contact', i.e., someone who provided an 'entry point' for us as researchers.

One of **Group B**'s pair of partners were often apart. The other three participants were friends studying at university. All group members were recruited from and acquainted with one another via a local sports club. A common interest (mountaineering) brought the researcher and participants together. Coordination and instruction was provided with all members of the group rather than via one 'primary contact'. With the exception of one member, **Group C** had attended school together and grown up alongside one

Group	A	B	C	D
Age range	24-50	Early 20s	Early 20s	Late 20s
Football fans?	N	N	Y	Y
Recruitment method	Friend-of-friend	Sports club	Press coverage	Friend-of-friend
Social configuration	Mother, daughters, niece (5f)	Two partners (2m, 2f)	Friendship group (5m)	Friendship group (3m)
	Friend of family (1f)	Friend (1f)	Cousin (1m)	Friend (1m)
Trial length	2 weeks	2 weeks	4 weeks	4 weeks

Table 1: Outline of trial groups

another for approximately ten years. The final member of this group was a cousin of one participant, and was less well-known to the larger group. This group had a primary contact who helped to distribute phones, instructions and so on. Recruited via press attention around the project, in many ways we found this produced a social distance with those participants. Lastly, and again with the exception of one member, **Group D** had grown up and attended school together. The group had know one another for approximately twenty years, except for one participant who instead knew the other members of the group via a football supporters' club.

Our data collection consisted of logged information from the trial itself (records of participants' interactions with the system), and interviews with the participants themselves at the end of the trial. During interviews, we reviewed with participants the content their group had produced. As part of this we used questions that had been developed before the trial, and questions developed over the course of the trial.

Results

Broadly, as with nearly every photo sharing system in the literature, the system was used to a *reasonable* degree, with 17 photos and 19 comments produced by each participant and roughly 1 photo or comment per day per participant. But what did the participants think about the trial itself and their own behaviour?

Demand characteristics

One key aspect of the users' descriptions of their own behaviour was what is known within the psychology literature as 'demand characteristics' [26]. Demand characteristics are where users shape or enhance their behaviour in a trial or experiment, in response to the imagined desires of the investigators. For example, users might increase their usage of a system if they assume that system usage is what the investigators are seeking, or deliberately ignore the system to reject the investigators involvement. This phenomenon is related to the well-known 'Hawthorne Effect' (while well-known, the original work that coined the term is somewhat problematic [19]).

In order to assess the importance and form of demand characteristics present in our trials, we asked participants to reflect upon how they felt the trial was run, what, if any, expectations they felt in taking part in the trial and what motivated them to engage in system use. Much of participants' motivation to use the system seemed to stem from a sense of obligation to us rather than their relationship to the system *per se*. Participants were typically eager to emphasise in interviews when they had been a 'good participant', often highlighting that they had been "us[ing] it everyday as in checking for other updates", "taking [it] to work and stuff every day", or "carrying it round all the time". Participants in the groups discussed how they attempted to assist us as experimenters, searching out ways to actually use the system: "I was just thinking what can I take photos of" or "trying to [...] think of things like on the two weeks [of the trial] ... of what to do". The desire to help us also extended to the way in which participants responded in interviews: "I was trying to make sure I had a good list [of suggestions] for you, good things for you". Participants would go as far as to account for any lack of use (as they saw it) in terms of a problem with *themselves* rather than the system ("I guess I just didn't make many comments, sorry I should I guess I should have"), whereas others suggested their less frequent use of it compared to the rest of their groups was due to *themselves* being "too boring", "[not] interesting enough". Some participants described their efforts to "show [some] willing" in spite of perhaps being "a bit of a technophobe" or even apologising for feeling "too old for it" - "[I] think oh I need to take a photograph because I was doing this the thing for you". We noted that participants treated these expectations towards 'making the system work' as obvious, normal and unremarkable ("obviously you need the data").

As part of this sense of obligation, participants also encouraged one another to use the system, and organised when they might use it together. One participant reported discussing amongst their group how they would "take [the phone] with [them] to [their] work and everyday use" so that they, as a group, "just use it when [they] could". Participants from another group described how they coordinated their use of the system such that "one of us would be at the football, and then you would text someone else to say I'm on FanPhoto tonight". Again such motivations for interacting with others, besides an interest in such interaction for its own sake, were accounted for in terms of how, as a group, they might ensure that they provided us with 'good data'. Participants also spent considerable effort interpreting the trial's purposes and the relevance of their own actions in that light. Often there would be talk of what they considered to be relevant or non-relevant activity, reflected upon the corresponding potential of activity to be 'good data' for the trial or

'irrelevant'. When interviewed, one participant, for instance, highlighted which photos and comments he considered "irrelevant" as opposed to those photos and comments that, as he put it, were "meaningful" and of being "of real consequence" (in this case, a section of a match report). Of course much of this was influenced by how we set up the trial and whether we described the system as a 'football' or general 'photo' system.

This concern for behaviour of relevance extended beyond our relationship to the participants, and also was an issue of what was relevant *between* participants themselves. In interviews participants described how they were attentive to relevance for their particular group. Thus, as one pointed out from our football fan groups, "I tried to use it more as a kind of football thing, if the football was on I'd make more of an effort", and as another mentioned, "I was more looking toward taking pictures that were relevant to an event". For Groups C and D, participants characterised importance in terms of football photos and this aspect also reinforced our presentation of the trial to participant groups. In other words, participants conducted their interaction with the system according to assumptions about the relevance of those actions to the purposes of the trial and the expectations of the experimenters. One member of Group D reported trying to keep his activity "within the guidelines" (presumably according to the way in which the system was presented at the start of the trial).

That said, participants did reflect upon their own role within the trial - for instance, in discussing "what [we] were looking for", a participant suggested he was "mostly kept in the dark" regarding the trial's details, and that he "supposed that's good really". This further highlights participants' normative attitudes towards trials, perhaps based upon previous experience of trials or common knowledge of experimental method. Participants often drew our attention towards sequences of activity that they presumed we would find useful (e.g., "this is the one I wanted to tell you about") and set aside those instances they considered less 'relevant'. As Gaver et al. points out in their own trial: "The continual engagement with the system appeared motivated as much by questions about our research agenda as by interest in what the system was saying" [15]. In one case two participants from Group C brought up in the interview a sequence in which one of them requested of the other that he update him on the progress of a game while he was attending a German language class. Similarly, in Group D, two participants both drew our attention to a sequence of photographs in which they and another participant captured a view from their positions within the stadium: "we were all at the same game but we were all in different bits of stadium … and obviously different tickets so it was quite good to like upload photos to say look this is where I am, where are you". In these instances, and throughout interviews, participants presented selections of their conduct to highlight as 'good activity'. Again, as before, these selections were driven by what participants interpreted as relevant to the purposes of the trial.

Lead participants and social relations

Alongside participants being oriented towards the trial, and each other's system use, in one group we found more formal and explicit orienting work. One participant in particular (in Group C) encouraged others in that group to engage in using the system, with the 'lead participant' (as we shall call him) making various demands on his group members. Other participants described with varying degrees how the lead participant was "out of everyone [...] very enthusiastic [...] he would kind of take a lot of photos and encourage us to take photos". In addition to encouraging others to generate content through creating content himself, he employed more direct strategies, such as asking another participant to take a photograph of his view of a stadium from his workplace (since he happened to work nearby), or emailing and text messaging others, say, to "get it out again for the game on Saturday".

This 'lead participant' also highlighted for the others why they had the system, as one participant mentioned "this is what you've got your iPhones for, this purpose". Although other participants in the group did engage in some encouraging activity, for the most part participants took their lead from this member of the group. The other groups (A, B and D) were characterised more by a sense of obligation to us as experimenters alone. As one participant reflected on the relations in his group (D), "I know the guys I wouldn't say I was really close to them but I know them and I've had some some kind of banter with them so there was there was an expectation there but to less of a degree".

More broadly, the social relations within the groups influenced the results of our various trials. As outlined earlier in our brief profile of each group, the trial groups were widely varied in terms of how they knew one another, for how long, their gender, their personal relationships with one another, how often and how much time they spent together, what their interests were and how those interests coincided with the interests of others. The relevance of social ties were formative in understanding the different patterns of interaction with and via the system. For instance, members of Group A were all equally close friends, and engaged in notably more extensive and mixed exchanges of comments than Group B. In Group B, we see far fewer exchanges due to the stratification of the group that did not occur in Group A, i.e., a less homogenous group which consisted of a couple and three close friends as opposed to a group of equally close friends. Between these two subgroups within Group B we would argue that the nature of being acquaintances influenced the reduced commenting that occurred. Group C consisted mostly of close friends who had known each other for a long time, and whose friendship was formed on the basis of common interest in football, although more recently the group had been less able to meet up for watching football together due

to lifestyle changes ("back when we were younger we would have watched midweek games together sort of without fail but now we're not meeting up to watch it as much"). The system potentially offered a method by which to return to 'old times' that some participants in Group C noted as spurring their usage.

The social relations of Group C also contributed to non-participation; one member of the group did not know others so well so his participation was correspondingly less, and less 'integrated' in the social interactions (via comments) of others. In particular, he supported a different, more obscure team to the majority of the other participants, contributing to his reduced interaction via the system: "I noticed that the [other team] fans were all chatting about the game and obviously I couldn't say anything about it cause I wasn't there … so I'd say it would be good to have if it would be different if there was a fellow [team] fan".

The role of trial design and framing

Another facet of our trial was how we presented the system to the different participant groups and our role as investigators. This formed an important orienting influence in the way that participants engaged with the system, with us as experimenters and with one another. Although the instructions for actually using the system itself were delivered to each group in an identical manner (via instruction sheets), the purposes and reasons for the trial and participants' engagement with us was discussed in a way that was sensitive to the different groups. Thus, for Groups A and B, we introduced it by explaining that we had designed the system as a simple photo sharing application, and were recruiting them in order to test the quality of the software and to work out how we might iteratively improve it. For the second two groups of users (C and D) we stated that the application had been designed in order to support groups of football supporters in sharing their interests in football matches with one another, alongside a recruitment process which had been focused around football. Although the formal instructions given concerning the system were identical, in practice the differences in recruitment and in how the system was verbally framed led to key differences in use.

By and large Groups C and D directed their use around footballing events and the vast majority of their photos concerned football in some way. Group C sustained their use throughout the trial period, and their production of comments and photos was heavily oriented around ongoing football match events. As we have seen, the conduct of 'lead participants' probably contributed to this. For Group D, usage rapidly declined after the 'main events' (two football games) were over, and participation in the generation of content was more markedly slanted towards certain participants (i.e., one participant in particular contributed the majority of photos to the group, both from a trip abroad and as part of sharing his collection of football shirts). There seemed to be an important relationship between the nature of the events and participant use of the system. As one participant commented, "[The system] is better when the [football] game's better". In this way the use of the system by Groups C and D was more dependent upon events and their relative quality, which itself was directed in part by the way that we introduced the groups to the application (i.e., as an application to support football experiences). This may explain also why Group D's activity declined rapidly midway through the trial - since they as a group participated in fewer events, and their use was more event-based, the system no longer was interesting to use. In comparison, Group C sustained their use through attending many more events.

Correspondingly with the different framing we provided, Groups A and B did not engage in event-based use. Instead, photo and comment activity was woven more into workplace and home activities, which are not event-based in their nature. Interestingly, participants in both Groups C and D noted in interviews that the application could easily be "used for anything" even though participants still organised the majority of their participation around sporting events (no doubt due to the nature of their social relations being based on common interest in football). As one participant put it, whenever there was an event he deemed relevant (namely, football) he would "make more of an effort" to capture something. In spite of presenting the system in a particular way, some participants noted that they were "not sure what [we] were looking for" and since they didn't know the purposes of the trial they were not clear if they had provided us with the data they assumed we desired.

LESSONS FOR TRIAL METHODS

We make these points not as criticisms as such of the use of trials - the very inspiration for research 'in the wild' is abandoning notions of purity or simple deterministic relationships between technology and use. This discussion highlights aspects of trials that we have noted in our earlier trials, yet seldom highlighted in the published presentation. Our goal rather is that by documenting realities of how trials unfold and are influenced by these factors (demand characteristics, lead participants and trial design) we can draw lessons for how to better conduct and present trial data, how we can exploit these features so to run better trials.

Demand characteristics: investigators as participants

As outlined above one source of behaviour for trial participants is is their interpretations of what would be seen as the 'right' behaviour for those running the trial. At first blush this might seem to present a challenge to those running trials - one response would be to accentuate the distance between participants and investigators, and to attempt to adopt more neutral stance with participants. Yet this would not eliminate demand characteristics, and participants would simply resort to their own preconceptions of 'typical trial behaviour'. Demand

characterises are not an example of bad methods, but are instead a fundamental part of what makes trials possible in the first place. Participation in a trial requires generosity in giving one's time and energy to taking part, particularly in trials that take place over an extended period. Even when incentives, such as payment, are given to participants this is seldom the only motivation (and if it is that can be problematic). Demand characteristics - the desire to produce something of value for those running the study - are thus fundamental to the success of trials in the first place and in many ways trials depend on demand characteristics for their feasibility.

Yet in descriptions of trials there is usually little written about how participants orientated to investigators, or how the investigators themselves saw the system being tested and the trial as it was run. What we are suggesting is the inclusion of '*investigators as participants*', seeing them as 'inside the loop', rather than controlling the trial from outside. By seeing investigators as participants in how we think through and write up trials, demand characteristics can be seen as a natural part of how trials are run. Much of what happens in a trial in this sense is an interaction between participants and investigators - one with particular obligations and requirements. On a practical level this encourages better documentation of these interactions, but also an explicit orientation to how these interactions shape users' behaviour. For example, in own trial the loaning of an expensive mobile phone for the extent of the trial acted as a considerable incentive to use the system. The 'gifting' of equipment was key in encouraging use, and thus providing usage and data for us to analyse. The arrangement of our data collection - with individuals from the groups interviewed individually also produced an individual accountability for use. As the trial developed, our own interests as investigators turned to how the relationships between participants, ourselves and how the trial was run. This in turn generated behaviour by the participants orientating to these interests. Our participants were thus very likely well-tuned to how their behaviour might be seen in terms of these relationships.

Indeed, demand characteristics can be exploited to an extent to encourage usage in trial settings. So, for example, participants could be given feedback as a trial developed on the observations that are being made, and a commentary passed on different forms of use. This would act as a way of encouraging involvement by participants in the trial, through building a reciprocal relationship. Even being explicit about what is 'expected' of participants could give participants the possibility to accept or reject those expectations explicitly. Questions about demand characteristics come into particular focus with the recent use of 'mass participation' trials, through new software distribution methods. While one might get relatively high numbers of downloads, motivating engaged use is still an open challenge here, partly due to the lack of demand characteristics amongst those who download trial software [24].

Lead participants: participants as investigators

A second methodological point concerns the role of participants themselves in trials. Frequently a subset of trial participants becomes key in how a trial is run and results are drawn. These participants - or even participant - engage with the technology and reflect on its use by themselves and others in a particularly insightful way, or alternatively work so as to encourage involvement by others who are involved in the trial. In our own trial although users gently encouraged one another to use the system, the presence of a particular highly enthusiastic 'leader' significantly drove interaction greatly in one groups. Further, the lead participant did considerable work in the interview to find interesting incidents, to offer suggestions as to what was interesting from the trial, and so on. This is not a rare event in trials in our experience, and we have frequently depended upon the insights or activities of lead users who either offer particularly astute assessments and reports of their own and others' behaviour, or go so far as to reflect upon and adjust their own practice as the trial progresses. For example, in [1] we noted how one player in particular extensively documented his own play (via video recordings), and encouraged other users to play the game more frequently.

The challenge this presents is that the use which is most interesting in terms of analysis is the use that is actually *least* typical. Yet analytically though even if only one participant uses a technology in a particularly interesting way, this atypicality is irrelevant - behaviour around technology is something that develops over time and participants are not a simple 'sample' of the greater population. Participants are being used as experts on their own activity, attempting to predict what might happen with a particular technology, to develop insight based on their use. The frequency of an observation has no relationship to insightfulness. More broadly, what we are suggesting is that *participants can be seen as investigators* themselves in the trial, with a move to acknowledging that it is through participants' own insights that the power of the trials can be focused. In some senses this echoes participatory design, yet here we are not arguing for users as designers but rather users as *analysts* of their own and others' practices.

In practical terms we see considerable potential in expanding the contribution that participants provide towards results, beyond being passive subjects. One example (drawing on the concept of lead users from technology marketing) is using blogs or the commentary of technology enthusiasts as a resource for allowing more lengthy review and discussion of technology by participants [22]. Gaver's work on cultural commentators [14] similarly explores the potential of participants as investigators for the design process. A related approach is narrative inquiry which has developed a set of methods

around interviews that go into a much more lengthy examination of individuals' life histories and perspective ('working in a three dimensional narrative inquiry space' as Clandinin and Connelly [8] put it). While these methods might seem excessive, what they do is move away from the notion of participants as interchangeable 'blank slates' and engage with users as diverse active participants in the studies we run.

Trial design: diversifying methods

A final challenge identified from our trial was in how the relationships between the participants, and our framing of the system at the start of the trial, led to very different types of use between the different participants. This echoes our argument that participants are not homogeneous, and that the particular characteristics and distinctiveness of participants can lead to very different trial events. Moreover this suggests that different trials, run in different ways, can end up with very different results even if the system that is being built is exactly the same (e.g. [1]). The informal details of what goes on in the trial (the personalities of those involved, the culture it takes place in, how the system is introduced to the users) all influence trials. This suggests the value of running trials with the same systems multiple times in multiple different ways as ways of getting corpuses of diverse results. One example of this has been the diversity of experiments around location tracking technologies, systems which frequently feature very similar functionality, but via diversity in the experimental form have produced a diversity of interesting results [29].

DISCUSSION

One response to our arguments here would be that these points are already well known to those who conduct trials. Certainly, to researchers intimately involved in the production of qualitative, interpretivist research, the deep involvement of researchers in the data collection and analysis are foundational points. As Mauthner puts it, 'we are the data' [23], i.e., researchers are intimately bound to its production. Yet how this plays out in trials - as examined in our study - has a number of implications for how we might develop trial methods further.

Rejecting reproducibility

A first point is the rejection of those who advocate more standardised approaches to technology trials. There are few that are explicit in this call; one exception is Whittaker, Terveen and Nardi [31] who argue for a standardised set of trial protocols, with the use of these to support replication of experiments multiple times by different researchers. This is an aim for a comparability across experiments that is currently impossible with the diversity of approaches taken. We would suggest though that even with the best efforts of researchers, the issues that we have identified above would make such reproducibility impossible. Participants in trials are not interchangeable but are instead individuals who will relate to trials in their own way. They have a diversity of social relationships with each other and with investigators - relationships which as we have seen have a direct impact on system use. It is thus critical not to underplay both the role that we as experimenters have in the research that we do, and the natural variability of trials. It is not just that the large natural variability of humans, or variations in trial procedures, that makes the standardisation of trials impossible. It is that this goal is itself misleading - social settings involving humans and technology contain far too much variability to be reproducible in any straightforward way. While Whittaker et al.'s argument is perhaps an extreme case, the desire for comparability and standardisation in trials is not unique to them.

Moving beyond success

A second point concerns the ways in which trials suffer from the prevalent normative orientation towards 'success'. Systems and their trials in are often written in such a way as to presents the system as a success with users. This is an understandable and natural tendency - on completing a difficult technical project one wants to validate the success of at least the technical work and the ideas behind it. It is even implicit in the term 'trial', i.e., a test of suitability. This leads to the presentation of results in papers orientated in subtle ways towards highlighting users' complimentary comments on a system, or (more bluntly) only highlighting the aspects of a system that seemed to work with users in the trial. Yet clearly if one takes demand characteristics seriously, compliments about a system cannot just be quoted verbatim and taken as evidence of its straightforward success. User *compliments* are in some ways to be expected within the framework of an interview. Even usage statistics are problematic in a similar way, and although informative, should not necessarily be taken as blunt indicators of 'success'. Moreover, whatever the seeming success of a trial, how a particular technology might fare in non-trial contexts can only be broadly ascertained from what happens in that constrained setting. Recently, Gaver et al. [15] have argued that we need to acknowledge and engage with understanding failures in trials. This is something we have much sympathy for - certainly there are aspects of many of the systems which we have trialled in the past that simply did not work with users. Yet this advice could come to exacerbate the problem in that it encourages analysis in the straightforward evaluative terms of 'good or bad'. This question is not only frequently unanswerable but can limit our understandings of what is going on with technology, since each element of a system can fare differently. What happens in a trial can be indicative of the concept, that particular instantiation or even just the form factor. We would argue that we do not need to discuss failure more in HCI, but rather to break free from the premature evaluation of technologies before understanding how they interact with users and practice.

Rewriting the methods section

More boldly we would argue for innovation in how methods sections are presented in trial papers. Changing the focus of reporting trials from offering up replicable results invites a set of changes in how we describe our trials, and in particular the form that the methods section takes. Methods sections in user trials papers are, broadly speaking, written in a 'standard' way. Often this includes statistical information outlining the make up of the particular user groups. These include age, sex and some other basic contextual information such as occupation, along with some brief notion of how the users are connected. Moreover, there is a strong normative orientation in writing methods sections to presenting results as predictable, replicable and comparable. Indeed, critiquing the methods of our own papers, their orientation has been almost entirely defensive rather than documenting the actual details and tribulations that trials practically, really face. The methods sections of papers are often repetitive and anodyne, skipping much of the details of how a trial actually proceeded so as to prevent provoking reviewers.

To truly embrace the distinctiveness of trials we propose that this additional context is extended much wider and documented in greater detail. Methods sections should be more explicit about the natural contingencies and events that happen while a trial is carried out. These are not signs of a 'bad trial', but are important details that lets us understand better the differing contexts of particular trials. What is needed are methods and results sections that allow us to interpret where sources of variation come from, and the different ways in which trials are planned yet transpire in quite different ways. One example of this is [25], where a system failure resulted in an important and key research finding. Obviously, one key reason behind the formulaic nature of methods sections is the reactions of reviewers. Expanding and enriching the methods section of papers perhaps most of all will require modifications in reviewers' approaches. Frequently, there is an attempt by reviewers to find the 'fatal flaw' in a methods section and this results in methods sections written in a highly defensive manner. While this might be applicable to positivist approaches, we would argue that it is inappropriate to trials and in the longer term has had a negative impact on how methods are reported.

Lastly, we would argue for much greater innovation in methods around trials, a break away from the assumption that trials should be as 'natural' as possible. While we are not arguing that 'anything goes' in terms of methods we have some sympathy for Feyerabend who put forward an argument that, as science changes, often innovation happens as much in the methods used (and what comes to be seen as a fact) as in the actual discoveries of scientists themselves [12]. Science, as Feyerabend argued, does not rely upon a single universal notion of what is truth and how we find it; the very 'rules of the game' change as science innovates and moves into studying new phenomena. Feyerabend recommended 'methodological anarchism' as a way of increasing innovation in science - and that scientists should seek to innovate much in terms of the methods they use and in how they attempt to prove their findings.

CONCLUSION

Our goal in this paper is to focus attention on one of the key user studies methods employed in HCI. We have drawn on a 'trial of trials' - a study of a well-known, predictable technology where we focused our attention on understanding some of the complexities of how users behave in trials, rather than on the system itself. We documented the many different sources of behaviour in trials: relationships between investigators and participants, relationships between participants themselves, and the nature of trial instructions. These sources of behaviour are often neglected in trials discussions. From this we argued for three key changes in trial methods: a move away from an orientation towards the 'success' of systems; innovations to how methods sections are presented that encompass and document the reality of trials as practiced; and a broader embrace of innovation in methods, one that fully acknowledges the role that the investigator plays in user trials. In general we call for participation in an ongoing process of innovation with regard to our methods. For this to happen we would argue that there is need for a significant shift to be made by reviewers as much as practitioners.

REFERENCES

[1] Barkhuus, L., et al. Picking Pockets on the Lawn: The Development of Tactics and Strategies in a Mobile Game. In *Proceedings of UbiComp 2005, Tokyo, Japan.,* Springer Verlag (2005).

[2] Barkhuus, L. and Rode, J. From Mice to Men: 24 years of Evaluation in CHI. *Extended Proceedings of CHI 2007 (alt.chi)*(2007).

[3] Bell, M., et al. Interweaving mobile games with everyday life. In *Proceedings of CHI 2006,* ACM Press (2006), 417-426.

[4] Benford, S., et al. Experiments in inhabited TV. In *Proceedings of CHI 98,* ACM Press (1998).

[5] Benford, S., et al. The Error of our Ways: The experience of Self-Reported Position in a Location-Based Game. In *Ubicomp 2004,* Springer (2004).

[6] Boehner, K., et al. How HCI interprets the probes. In *Proceedings of CHI 2007* (2007) ACM.

[7] Carter, S., et al. Exiting the cleanroom: On ecological validity and ubiquitous computing. *Human Computer Interaction*, 23, 1 (2008), 47-99.

[8] Clandinin, D. and Connelly, F. *Narrative inquiry*. Jossey-Bass San Francisco, 2000.

[9] Crabtree, A. Taking technomethodology seriously: hybrid change in the ethnomethodology-design relationship. *Eur. J. Inf. Syst.*, 13, 3 (2004), 195-209.

[10] Davies, N. Proof-of-concept demonstrators and other evils of application-led research. *Proceedings of UbiApp Workshop, Ubicomp 2005*(2005).

[11] Dourish, P. and Adler, A. Your place or mine? Learning from long-term use of audio-video communication. *Computer Supported Cooperative Work (CSCW)*, 5, 1 (1996), 33-62.

[12] Feyerabend, P. *Against method*. Verso Books, 1993.

[13] Gaver, B., et al. Design: Cultural probes. *interactions*, 6, 1 (1999).

[14] Gaver, W. Cultural commentators: Non-native interpretations as resources for polyphonic assessment. *International journal of human-computer studies*, 65, 4 (2007), 292-305.

[15] Gaver, W., et al. Anatomy of a failure: how we knew when our design went wrong, and what we learned from it. *Proceedings of CHI 2009*(2009).

[16] Harper, R. H. R., et al. Locating systems at work. *Interacting With Computers*, 4, 3 (1992), 343-363.

[17] Hutchinson, H., et al. Technology probes: inspiring design for and with families. *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*(2003).

[18] Intille, S., et al. A living laboratory for the design and evaluation of ubiquitous computing technologies. In *Proceedings of CHI 2005* (2005),1941-1944 ACM New York, NY, USA.

[19] Jones, S. Was there a Hawthorne effect? *American Journal of Sociology*(1992), 451-468.

[20] Kjeldskov, J., et al. Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In *Proceedings of MobileHCI 2004* (2004),61-73.

[21] Lee, M. L. and Dey, A. K. Lifelogging memory appliance for people with episodic memory impairment. In *Proceedings of the 10th international conference on Ubiquitous computing* (2008),44-53 ACM.

[22] Ljungblad, S. and Holmquist, L. *Transfer scenarios: grounding innovation with marginal practices*. ACM, City, 2007.

[23] Mauthner, N., et al. The data are out there, or are they? Implications for archiving and revisiting qualitative data. *Sociology*, 32, 4 (1998), 733-745.

[24] McMillan, D., et al. Further into the wild: Running worldwide trials of mobile systems. *Pervasive Computing*(2010), 210-227.

[25] Pousman, Z., et al. Living with tableau machine: a longitudinal investigation of a curious domestic intelligence. In *Proceedings of Ubicomp 2008* (2008), 370-379 ACM.

[26] Robinson, M. Design for unanticipated use. In *Proceedings of ECSCW 1993* (1993),187-202 Kluwer Academic Publishers.

[27] Rogers, Y., et al. Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. In *Proceedings of Ubicomp 2007* (2007).

[28] Root, R. W. Design of a multi-media vehicle for social browsing. In *Proceedings of CSCW 1988* (1988) ACM Press.

[29] Smith, I. E., et al. Social Disclosure of Place: From Location Technology to Communication Practices. In *Pervasive 2005* (2005), 134-151.

[30] Tolmie, P. and Crabtree, A. Deploying research technology in the home. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (2008),639-648 ACM.

[31] Whittaker, S., et al. Let's stop pushing the envelope and start addressing it: a reference task agenda for HCI. *Hum.-Comput. Interact.*, 15, 2 (2000), 75-106.