What Do Lab-based User Studies Tell Us About In-the-Wild Behavior? Insights from a Study of Museum Interactives

Eva Hornecker, Emma Nicol
Dept. of CIS, University of Strathclyde, Glasgow G11XH, UK
eva@ehornecker.de, emma.nicol@strath.ac.uk

ABSTRACT
We contribute to an understanding of how well lab-based user studies can help us to anticipate how a system will be used in 'the wild'. We analyze and compare data from lab-based user studies of prototype museum installations and the subsequent deployment of these systems in a museum. While the user study was successful in identifying usability issues, social behavior patterns in the museum, in particular between caregivers and children, differed in several aspects between the settings. Our analysis highlights influences on usage and behavior patterns: the physical and structural setup, the user study creating a focused activity, and the demand characteristics of a user study.

Author Keywords
Museum, CSCW, family, user study, social behavior

ACM Classification Keywords
H5.2. User Interfaces: Evaluation/methodology.

INTRODUCTION
There has been a long discussion within HCI about the role of lab studies and whether these are sufficient for uncovering usability issues that are influenced by the context of use (e.g. movement, distractions) [10, 22, 28]. More recently the need for in-situ studies of UbiComp technologies has been emphasized [4, 29]. Field trials are deemed indispensible to assess how UbiComp technologies fit into people's lives and how well they work under real-world conditions [3]. Understanding the limitations of lab-studies is not just relevant for research, but a very practical question for system designers and evaluators. These need to decide where and when to test, and sometimes may have no alternative to lab-based user studies for practical reasons.

Here we contribute to a refined understanding of the difficulties of extrapolating from lab-based user studies of system prototypes on future 'in-the-wild' use. In particular, we are interested in how closely a lab-based study can emulate the *social use* situation. It is commonplace that context matters, but it is important to know how and why.

DIS 2012, June 11-15, 2012, Newcastle, UK.

We present a concrete case study in the context of visitor interaction with museum installations. We were contracted to run user studies of early prototypes of several museum installations for a new museum. Based on our familiarity with the museum domain [13, 15], we made an effort to generate a semi-realistic setup, in particular by inviting families as participants. Despite all efforts, we felt that this left several questions open and that participants were subtly influenced in their perceptions and behaviors (see [14]). We therefore followed the research up with an observational study in the re-opened museum.

While we found that most usability issues observed in the user study (unless remedied) were replicated in the museum and no unforeseen usability issues emerged, the picture regarding use and social interaction patterns was diverse. Our study setting did not convey how much systems would be used and how. In particular, there were marked differences in family interactions and parental behavior, for instance in the amount of parental supervision, scaffolding and 'educational talk', which in turn influenced children's behavior. These effects are not uniform across all installations, and were influenced by factors such as physical placement and setup. Our analysis highlights the effect of user studies creating a focused activity, and their demand characteristics.

BACKGROUND AND RELATED WORK
The methodological challenge of developing appropriate research methods for understanding the use of new technologies became particularly obvious for mobile computing, where user mobility, environmental conditions, and unpredictability of use provide challenges both for observing users and the design of lab studies [10, 21, 22]. Several research teams have developed strategies for more ecologically valid experiments, emulating relevant aspects of the use context, running quasi-experiments in real-world settings or situated evaluations [12, 25]. While traditional usability testing might not always be feasible, for example, if novel technology or designs mean that new use practices yet have to evolve, or the technology is not mature [8, 24], new approaches are emerging to e.g. explore UbiComp in the wild through interventionist studies or by designing in the wild [5, 29]. Yet, despite the need for methodological innovation and reflection, a recent survey of CHI papers on system evaluation [2] revealed that the number of papers about evaluation methods as a topic in itself is in decline.

The debate about the value of field trials versus lab studies is still on, of where *exactly* field studies are 'worth the has-

sle' [22, 28], what their added value is, and how far lab studies can predict real-world usage. Many researchers consider real-world evaluations essential for CSCW systems [26], which are highly influenced by social context, and UbiComp technologies, where these may reveal usability issues and group interaction patterns that do not arise in the lab [23, 28, 2]. Field experiments enable identification of factors affecting user behavior [25]. But to make matters even more complex, even field trials might not produce the same behavior as a deployment [3]. Challenges arise due to their by-invitation characteristic, which often has participants adjust their behavior and responses to the perceived expectations of researchers, feeling obliged to "be a 'good' participant" – or they might attempt to negate expectations [3]. This demand characteristic or facilitation bias is well known [6, 33] in usability testing – users interpret the social situation and want to 'do it right'.

Our study adds to this discussion through a systematic analysis of data from lab-based user studies and the subsequent deployment of (improved versions of) the same systems. It provides a deeper insight into the effect of demand characteristics in this context, focusing on how social behaviors, in particular between caregivers and children, differ between these settings. The systems we tested seem a fairly well understood and established genre, compared to radically novel UbiComp systems, but nevertheless our study reveals several factors affecting behavior in the wild.

Application Context: Museums

Over the past years, museums have served as a popular domain to investigate user interaction with novel technologies and for experimental system deployments [11, 13, 15, 31, 32]. At the same time, visitor studies research has begun to employ a wider range of qualitative methods to investigate the visitor experience, highlighting the sociality of museum visits [11, 20, 30, 34], and investigating what makes installations engaging [1, 9, 13, 16]. The museum context differs in many respects from other domains, influencing suitability of evaluation methods. Visitors want to be entertained and educated; their aims are highly personal, and change with what a system offers [7]. There is thus no 'task' as such. With many distractions, users quickly dismiss an installation that is not immediately satisfying [1]. User testing of installations thus needs to investigate usability and enjoyment. Moreover, the museum situation is inherently social – visitors come to spend time with family or friends, hoping the visit to be a memorable shared experience, and may also interact with strangers [11, 15, 20, 30, 34]. Adult-child interactions resemble those found when teachers and or parents engage with young children in technological play [26]. In this sense, museum installations are akin to party games, which are often playtested in living-room-like labs to study how players share control and communicate [18].

There is hardly any literature on the evaluation of (non-research project) early prototypes of museum installation. Museums tend to work on tight budgets, and development is often contracted to SMEs with no budget for in-house research and evaluation. Arguably the best strategy for user testing is to place a work-in-progress exhibit on the museum floor to observe visitors, systematically changing its features over time [cf. 16]. Yet this requires a fully functioning and robust system that can be used without staff support in a very chaotic environment. Thus, a "space of intermediate authenticity", giving a sense of how groups will react [9], might be preferable in many cases.

TWO SUBSEQUENT STUDIES

In spring 2010 we were commissioned by the National Trust Scotland (NTS) to conduct formative user studies of early prototypes of interactive installations for the newly rebuilt Robert Burns Birthplace museum (RBB) in Alloway, Scotland. Burns is lived heritage in Scotland. The new museum displays objects for veneration (original letters and objects), but in multiple ways also invites curiosity and playful interaction. It takes efforts to be family-friendly, and includes a number of interactive multimedia as well as non-digital interactive stations. The atmosphere is playful, with children running around, and loud music. The main visitor groups are young families and older adults.

Most of the installations tested take the form of mini-games and are aimed at engaging children with the themes of the museum: Burns' life, his poetry and the era he lived in. As the museum was being rebuilt at the time, we could not test in-situ. For the user studies we wanted to avoid a sterile lab situation, and to observe the natural social dynamics that evolve around installations. In particular we were interested in how socially scalable [32] these are to larger groups. The ability to entertain a family might be essential for a game's success in a museum. The recent museum studies literature highlights the role of family and parent-child interactions: parents explain, point things out, and ask questions [20, 30]. Moreover, one of the installations, an interactive table, was designed explicitly for cooperative gaming.

Therefore, to assess usability and fun of play, we enlisted young families and mature adult groups for the evaluation sessions, testing each installation with at least 10 separate groups. We recreated the installation setup to our best knowledge and ability. Over summer 2010 we successively received the prototypes, and ran each sub-study within a 2-week timeframe. Each sub-study resulted in a report on usability issues, participants' likes and dislikes, and detailed suggestions for improvement based on observation, post-session discussions and questionnaire responses. In many ways, our role was that of usability consultants to the NTS project. The prototypes had not undergone any prior user testing, the graphics were not detailed out yet, and the production system hardware was not available at this point. In [14] we describe the study approach in detail and discuss emerging questions about how the study setup influenced participant responses. The following year, we conducted an observational study of visitor interaction with the installations previously user-tested in the re-opened museum.

Figure 1. Test version of PhotoFit, with head selected so far. Final Spooky Stories Game with transparency on top of scenery showing the village. Prototype of Poetry Game. Mouse running below text and score (close-up in the inset).

The installations

The installations can best be described as 'mini-games'. They are part of a 'show not tell' interpretation strategy, which helps visitors enjoy Burns' heritage in an imaginative and playful way. We here focus on three touchscreen-based (single-touch) installations, aimed at children and teenagers.

PhotoFit (fig. 1 left) invites players to construct a photofit type image of Burns by selecting different combinations of eyes, mouth, nose, ears and hairstyle from available features. For each feature, a set of quotes from Burns' contemporaries on his appearance are shown on screen. The created face is then depicted on a shortbread tin.

Spooky Stories illustrates the 'Tam O'Shanter' poem. A villager leaves the pub on a dark night, encounters devilish creatures, flees, and rides over a bridge, where the creatures cannot follow. The game shows the village, and a virtual transparency of spooky figures that can be pressed out onto the scene and moved around within it. Once completed, an animation shows Tam walking through the village.

The *Poetry game* asks players to 'keep in time with rhythm and rhyme' by tapping a button. It plays the beginning of 'To a mouse' and Tam O'Shanter, while the lines move over the screen, with pieces of cheese placed over words at moments of emphasis (fig. 1). Tapping the button has a mouse that runs beneath the text jump at the cheese. After completion, the poem is shown on-screen with a score.

We furthermore tested the *Burns Supper Table*. This is a multiplayer game in five rounds, based loosely on the Burns supper tradition and played around a top-projected interactive table using *physical* buttons. The minigames ask players, for example, to stab haggis moving around the table. The button press makes the image of a knife poke out.

The final versions of the installations have more sophisticated graphics (only Spooky Stories had the final graphics), extended or refined animations, revised instructions, are improved in usability, and partially extended or revised, for example with more images to pick from in PhotoFit.

A Semi-Realistic User Study of Prototypes

Since technical and contextual setup influence usability and social experience of play [17, 32], we emulated the final setup as best possible. The games were tested on a touchscreen approximating the final installations' size, (20"), height and angle (fig. 2). We further built a table of comparable size and setup to the Burns Supper Table to assess how the game was played in a group, using an Arduino controller and top-projection.

Given the museum is visited largely by young families and adult groups, it is vital that the installations are enjoyable for groups and scale to different group sizes [13, 32]. Each game was evaluated with at least eight families with children of different ages, and two pairs of older adults. For about half the families, two or more children attended, often siblings, and the other half were adult-child pairs. We also asked families to bring a friend with a child, resulting in a few larger groups. We ran most studies within a cordoned-off area in the library of a local museum (fig. 2), to invoke the museum context. Due to the complex setup required for the Burns Supper Table, it was tested in our University lab.

Each group was welcomed, handed consent forms, and then invited 'to go and play' with the game, following a limited instruction procedure employed for user studies of games [18]. After the game play, parents and older children filled out a questionnaire while children drew pictures. Sessions ended with a short open-ended group interview. Each session lasted 40 to 60 minutes. All sessions were recorded on video (unfortunately, we had a total loss of video for Spooky Stories) and we took detailed observational notes. For about half of the sessions two observers were present.

The In-Situ Observational Study

When the museum reopened, we started the second phase of our research. We began with an open-ended (video supported) observation in the field. The observations were guided by questions left unanswered by the prior user study, and the question whether similar interaction patterns would occur 'in the wild', but remained open to emerging issues. For example, the *ShadowPortrait* installation, which had not been part of the user-tests, was placed directly next to PhotoFit and used a lot, influencing how the latter was interacted with. We thus partially extended our observation. Analysis is based largely on qualitative coding, while drawing upon principles of interaction analysis [19].

One researcher spent about 20 hours in the museum observing and video recording over the course of 6 days, mostly at weekends or school holidays. Another researcher also spent one day observing and taking field notes. Observation focused on the four installations we had user-tested. Visitors

were informed about the research via a poster at the ticket desk. Video recording was complemented with field notes. Observation summaries, emerging research questions and hypotheses were written up on the return of each visit.

For video recording, we resorted largely to using a handheld camera to provide a good view and audio of installations (this is a very noisy environment). This gave the flexibility to occasionally follow a group from one game to another. This strategy resulted in large number of clips of individuals or groups at installations. In addition, we installed a high-res camera in distant view of the PoetryGame for an hour. The interaction patterns and durations documented with the latter (as well as documented in our observational fieldnotes) are similar to those documented with the (possibly more obtrusive) handheld camera.

Data analysis

For this paper, we focus analysis on three installations. The open-ended observation indicated a strong discrepancy of behavior between user study and museum for PhotoFit, while the PoetryGame seemed to evoke similar group interaction. Whereas PhotoFit seemed less successful in the museum than the user study, the Poetry Game was successful in the museum, but appeared to be negatively effected by its setup. The third installation we focus on is Spooky Stories.

Observational notes and memos pointed to a range of issues for deeper analysis. For each installation, we transcribed the available video from the lab-based user study and the museum. Observational notes added detail not captured by the camera. When transcribing videos from the two settings, strong differences in family interactions became evident, and we decided to focus analysis on the issues described in this paper. Coded categories emerged out of analysis of transcripts, identifying recurrent behaviors or patterns. We began to define categories that were grounded in and developed iteratively from the data (with different sub-codes for different installations). We then systematically categorized and coded the available video data / transcripts (e.g. identifying instances of parents telling children what to do, to read off the screen, explaining etc.).

The following data was collected and analyzed (see table 1). PhotoFit and PoetryGame were each tested by two mature adult pairs and 8 families, with altogether 12 respectively 13 children. We analyzed about an hour of video clips of PhotoFit being played in the museum, and 1 hour of clips of the PoetryGame, plus an hour continuous footage captured from a fixed camera. While the fixed camera was unable to capture conversation and often people's bodies obstructed the viewpoint, it provided further insights. For Spooky Stories, we present a preliminary analysis of a random sample of 11 clips of museum visitors.

Figure 2. Study setup for touchscreen prototypes in museum library (top). Video recording view (video still)

GENERAL FINDINGS

We found that usability issues observed in the wild were largely identical to those in our semi-realistic user study, having been successfully 'predicted', while others had been resolved due to our advice. Since these issues are specific to the games and not relevant for the focus of this paper, we will omit describing them. In the following, we first discuss limitations of our user study [14] that partially motivated our in-situ study. We summarize findings regarding user behaviour, highlighting differences between lab and museum. Finally, we revisit these in detail along themes.

Limitations of the Semi-Realistic Setup

The user studies could not replicate the full social dynamics of a museum. Families were invited to participate, one at a time, and stayed together with nothing competing for attention. We could not assess large group social scalability [32] and whether strangers would play together. Moreover, users arrive continuously at museum exhibits [32] and might leave midway through an activity. Furthermore, people, even within a group, tend to come and go – interaction is more buffet than dinner-table-style [23, 29]. This can be an issue if a design assumes constant configurations.

	User Study (Lab)			Museum (In-the-Wild)				
	Adult groups	Families	Video	Adult groups	Families	Child groups	Total	Video
PhotoFit	2	8 (overall 12 children 3 - 13 yrs. mean age = 7.4)	1.3 hrs	8	18	2	28	1.09hrs (clips)
Poetry	2	8 (overall 13 children 4-11 yrs. mean age =7.5)	1.45 hrs	5	17	3	25	1h (clips)
				6	5	3	14	1h(continuous)
Spooky Stories	No video data			/	10	1	11	0.55h (clips)

Table 1. Overview of video data (number and type of groups, age range, overall length of video collected and analyzed)

In addition, the setup seemed to subtly affect the responses elicited in interviews and questionnaires [cf. 3], users interpreting the installations as stand-alone systems. They compared them with home video games, and made suggestions for extension or improvement that would not translate well to a museum context (increasing game-length and potentially frustrating small children). A contributing factor might be that the prototypes ran on a portable PC. Adults furthermore often found it hard to connect the games with Burns or his poems. Similarly, they often commented on the visual design as being outdated and boring compared with Nintendo DS©, despite of children enjoying the games. This suggests that participants were unable to imagine the installations as part of a museum visit, in spite of being in a museum (library), and the instruction to imagine encountering the installations in the new Burns museum. We anticipated further influences of the by-invitation status of user sessions [3, 6, 33], and it seemed likely that parents would not supervise children as closely in the museum, resulting in a wider variety of group behaviors. Altogether, despite of a semi-realistic setup, we concluded [14] that our user study nevertheless had many features of a lab-based study.

An Overview of Differences per Installation

Considering its simplicity, ***PhotoFit*** was surprisingly successful in the user study. Play patterns varied, with children playing rounds, in turns, collaborating or fighting over turns, and parents scaffolding and directing. Adults closely engaged with children. 3 of 8 family groups had to be asked to stop playing when running out of time for the session. The number of plays per family differed between 2 and 10. Families played on average 4.7 times, for up to 16 minutes, with each child involved in 4 rounds of play, and some for 10 times. Two children even asked to take the game home.

Observation in the museum showed that this popularity was not matched in-the-wild, and data analysis revealed strong differences in behavior compared to the user study. As we will discuss later, the vicinity of *ShadowPortrait* (an installation we had not been given to user test) influenced these. We observed that PhotoFit was frequently ignored and the game rarely repeated. Data shows an exponential decay curve for the number of plays per family group (see figure 3) (M = 1.8 per group). Not all children participated, and on average, each child played once. Adults tended to step back and observe, and often disengaged. Furthermore, people pondered little on their choices – reflected in the mean time

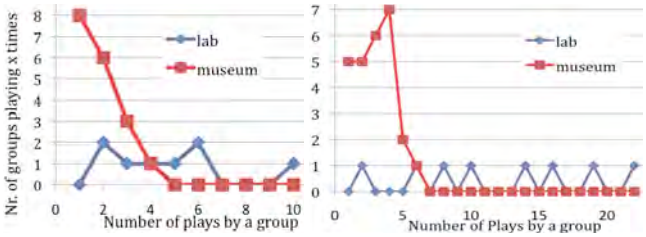


Figure 3. Frequency graphs for number of plays per group for PhotoFit (left) and PoetryGame differ markedly between user study and museum (note: different numbers of groups).

for one 'round' (creating a face) in the museum being 1-2 minutes, and 2-3 minutes in the user study. In the user study, the number of attempts to emulate Burns (or a realistic face) (12+3) equaled that of purely 'fun faces' (17), which were usually only done *after* two plays. In the museum only one group attempted to recreate Burns' looks (albeit this was easier now), compared to 24 'fun faces'.

With the ***Poetry Game***, museum interactions were more similar to the user study, albeit with less parental activity. The game's fast pace resulted in a pattern of children playing rounds (switching after poems) in both settings, and seemed to restrict opportunity for educational talk to in-between rounds of play, when the poem is displayed. Half of the user sessions had to be cut off when time ran out, and each child played on average 6 times, with up to 22 plays per group, and up to 13 per child. In contrast, 1 hour data from the fixed camera in the museum shows children only playing at most 4 times, with a tendency to repeat, on average twice (M = 2.45 if counting only those children that started play). Data of 17 families captured by the handheld camera indicates even lower means of 1.6 (for all children present) or 1.9 plays (only for children that began play) and a mean of 2.7 plays per group. Adults tended to only play once or twice. Only two adults in family groups played in the museum, whereas 6 of 8 did so in the user study. Solitary or groups of adults were more likely to play, and we informally observed some older adults playing.

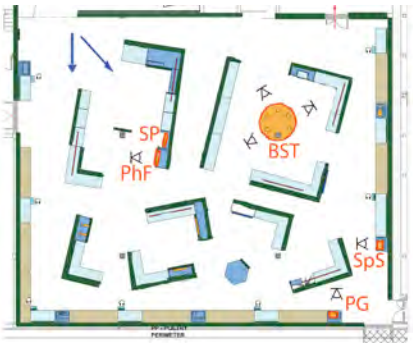
With ***Spooky Stories***, interaction between children largely mirrored that observed in the user study. Children predominantly played together (occasionally fighting for control) and sometimes in rounds. As we will show in the following sections, levels of adult scaffoldng in the museum were higher then for the other installations. The museum setting offered resources for educational talk from nearby displays, and many groups recognized the scenery. This is in contrast to the user study, where about half of adult participants expressed doubt users would recognize the poem.

For the ***Burns Supper table game*** we have not yet completed a full data analysis. Overall, in-person observation indicates that visitor interactions resemble the user study. Parents often join in the play, and children explain what to do. This shows that our setup using a physical table and similar interaction mechanisms generated realistic within-group dynamics. Overall, the table works as anticipated, and scales well. Strangers often play together, albeit rarely talk with each other. We often saw groups of up to 14 people from preschool to 60+ around the table. Furthermore transitions between groups appeared unproblematic, with a constant coming and going or interleaving of groups.

FINDINGS ALONG THEMES

The Influence of Physical Setup and Location

We now discuss how the physical installation setup [cf. 32] affected usage patterns. Figure 4 shows the museum floorplan with the installations highlighted in orange letters.



PhF = PhotoFit

SP = ShadowPortrait

BST = Burns Supper Table

SpS = Spooky Stories

PG = Poetry Game

Figure 4. Museum floorplan, installations marked in orange. Dark green is high, solid walls, seating is brown, display cabinets are light blue, non-digital interactives dark blue.

Other Installations in Direct Vicinity

Having Photofit with another installation right next to it resulted in different usage patterns than in the user study. At the *ShadowPortrait,* one can take a silhouette image of oneself (which appears in a gilded frame), email it, and view a gallery of past portraits. The portraits often constitute a group achievement, and people help, instruct, and physically move each other, comment and joke. Once somebody started, the entire group usually took portraits.

Especially larger groups tended to move back and forth between the two installations. PhotoFit often became a side-activity for a group member while others took portraits. The group's attention tended to be on the latter, and PhotoFit players kept an eye out for the final portrait (fig. 5 bottom row). They then tended to lean over towards ShadowPortrait, keeping a hand on PhotoFit screen's rim, marking it as occupied (fig. 5 bottom row, middle). PhotoFit here benefits from being interruptible and not dictating a pace. The divided attention of groups is reflected in children calling out *"look what I made"* to make (grand)parents look over for the produced face or shortbread tin. PhotoFit thus partially profits from the vicinity of ShadowPortrait (used as a filler while waiting), but at the same time suffers from the competition, as observing taking portraits is more engaging.

Location Location Location…

The Poetry Game was in a corner outside of the main passageways, which almost felt like a dark alley (fig. 5 top row). This affected how often it was used, with long pauses between sequences of play. Many groups who did not enter the alley did not notice it. Yet, once started, a group on average played 2.7 rounds (compare for PhotoFit: M = 1.8 per group). Then, other children might wait in the vicinity for it to be free, resulting in quick handovers. The setup, with nothing else to view or do apart from a bench for resting, moreover seemed to result in parents being somewhat less patient then elsewhere. The proportion of initiative to leave coming from adults was similar (~ half) across all analyzed installations. But at the Poetry Game adults were more explicit, tapping the child, saying *"lets go"*, one almost dragged it away (cf fig. 6 bottom right), and children resisted more, ignoring the prompt or protesting *"just once"*.

Other installations were located more centrally, letting parents engage with nearby exhibits while children remained in sight. PhotoFit was inspected and started frequently (even if many did not follow through). The interactive table drew many observers. Placed at a major hub on the museum floor, people saw it repeatedly from different angles and spent considerable time in this area, making it easy to 'mull around', waiting for a chance to play. Spooky Stories was placed less centrally, but within view of main passages, and its usage rate is between the other touchscreen games. Here, adults moved further along the row of displays, and appeared to not feel anxious to let children play on their own.

Height, Accessibility and Comfort

Finally, the height of installation placement affects use. Even very young children (2 years) enjoyed PhotoFit in the user study (assisted by adults). In the museum, the screen was flat to the wall and almost unreachable for toddlers. The two other touchscreen games were set up at an angle and height similar to the user study, ideal for small children. At the PoetryGame the button for beating the poem's rhythm is *under* the screen. Most children (60%) knelt or sat down, whereas most adults stood and bent over (62%). Spooky Stories was easier to play standing, but also had adults bend over. This is likely uncomfortable, and may cause the small number and length of plays by adults.

The User Study: A Focused Activity

Conducting user tests with one installation at a time (we needed to ensure that each game received detailed feedback), clearly constituted a lab-style setting, creating a focused activity. This is reflected in the high number of plays (see figure 3), in particular for PhotoFit, and the average length of sessions discussed earlier. Even though it was used often, visitors in the museum appeared much less enthusiastic about PhotoFit, often did not finish, and rarely repeated it. In contrast to PhotoFit, children often wanted to repeat Spooky Stories in the museum, which is more consistent with its lab-style evaluation. There seems no linear relationship between apparent attractiveness in a lab-based study and in real world use. Also, for the memory minigame on the Burns Supper Table, the first thing adults tended to utter in the lab setting was: *"I'm bad at this"*. Adults' reaction might, again, be due to them wanting to do their best and feeling they need to apologize for performing badly. This indicates how the user study generated a focused activity and clearly marked situation.

On top of the by-invitation characteristic [3, 33], the lab setting has little distraction, and participants have made time to attend. In a museum, the games compete with other activities and parents often want to move on. In half of the museum data, the initiative to leave came from adults, but also, other children in the group often ran off. The more distributed and less focused interaction with parts of a family attending to nearby installations or displays often had children call their parent to see the outcome of their efforts

(for the PhotoFit face or the animation at Spooky Stories), which in the user study setting was not necessary.

Adult-Child Interactions and Group Behavior

In both study settings we found that adults *scaffolded* children, *facilitated* interaction between siblings, and engaged in *educational talk* [cf. 27]. Scaffolding denotes helpful guidance and assistance, explaining or demonstrating how to interact, reading out instructions, prompting and guiding children. Adults further give emotional support, motivate, praise or alleviate frustration [27]. They often add context, point out things to notice, and engage a child in conversations that relate the current object to previous experiences [20, 30]. We refer to this as *educational talk*. Moreover, parents *facilitate* interaction between children to minimize conflict and ensure that all get their share. All of this occurred in both settings, albeit to different measures.

Figure 5. Location, location… Top: PoetryGame at end of a 'dark alley'. Spooky Stories next to a display illustrating the Tam O'Shanter Story, as seen from the interactive table. Bottom: ShadowPortrait in direct vicinity of PhotoFit - divided attention of groups moving between the two. The interactive table is placed in a central area.

A systematic analysis of adult-child interactions reveals further evidence of 'demand characteristics' [3, 6, 33] affecting adult behavior. These interpret the user study as a social situation where they want to 1) behave like a good parent and 2) help to make sure the study is successful. As expected, children were not always supervised in the museum. While it was rare for them to play alone (ca. 15%), in 30-50% of cases adults were not continuously present, either arriving well after the start of play or leaving early. The levels of educational talk, scaffolding, and emotional support were much lower in the museum. Differences were most striking for PhotoFit. Interestingly, the effect is not uniform, being less strong for the PoetryGame, and Spooky Stories saw relatively high levels of educational talk.

Keeping children on track

In the user study, parents were clearly attempting to orient children to the goals of the games. This indicates that they felt responsible to make sure children use the system correctly, similar to user attitude effects reported in field trials [3]. With PhotoFit, in 6 of 8 families (75%), an adult decidedly and repeatedly oriented children to recreate Burns' looks, read quotations of Burns contemporaries, and select matching features. The following are representative examples: *A girl selects an eye. Mom nods and points at a quotation about Burns' dark fiery eyes. The girl reads it out. Mom points again: "so if you see this, are you happy with that eye?" Girl: "Yes, I like that".* In another family, *the son selects a hairstyle. His mother comments: "I think that's none like him" and makes the son rethink his choice.*

In the museum, 14 out of 18 families never tried to orient children to select suitable facial features, two did to some extent, and two did gently but then gave up. The difference in attitude is reflected in the numbers of 'fun faces' reported earlier, but notably adults also mainly created 'fun faces' in the museum. For the PoetryGame, 2 of 8 families in the lab pointed out or reminded of the game's goal: *"you have to keep the rhythm"*. In the museum only 3 out of 17 families reminded children *"you are meant to be in time"* (even though many just pounded the button continuously).

Levels of Educational Talk

The setting did not uniformly influence levels of scaffolding and educational talk. While in general, there was less close supervision and less educational talk in the museum, it seemed to offer additional resources for educational talk at Spooky Stories. It should be noted that parents tended to be quite aware of their educational interventions in the user study (the questionnaire asked what role they tried to take).

In 6 of 8 cases, adults in the PhotoFit user study interacted closely with children. Only one mother did not engage in educational talk. Seven adults either extensively read out quotations about Burns' looks or prompted children to do so (*"what does it say here", "you read that one"*) pointing at the screen (see fig. 6 top left). Often they corrected or asked to re-read. Three went on to ask and explain what certain phrases mean ('hair without powder'), mimicking a ponytail. Adults often reviewed choices and asked whether the face created matched descriptions: *"which one is best?", "it says strongly defined nasal bridge – I think that one is best"*. In the museum, the amount of educational talk was minimal. Only 2 adults in 18 families pointed at and read out short sections of quotes. None of the adults observed ever asked if a child understood the quotations. Any reference to Burns tended to be jokes 'being not like him'.

With the PoetryGame, only one family in the user study did not engage in any educational talk. From the remaining seven, two asked the child to read out the poem (displayed

after play), four (half of families) pointed out its title and/or at the lines. Four asked whether a child knew the poem (*Father: "You know that poem?" – "We did it at school", "do you know what the story is about?"*), or talked about and explained the story (*"It is a man's name, who gets chased – the witches chase Tam O'Shanter"*). Moreover, four pointed out Scots' words and/or translated *("fou means drunk", "notice, some of it isn't spelt properly"*). Yet overall, educational talk was less frequent than with PhotoFit, and there was less insistence on 'doing it right'. In the museum, people rarely read out the poems and none discussed them. On the other hand, it was common for children and adults to mumble along to the audio, and other children to cling to the speakers to listen. 12 of the 17 families did not engage in any educational talk and merely observed. In three cases a section of the poem was read off the screen.

Our observational notes from the Spooky Stories user study indicate that almost all families engaged in educational talk, discussing where to place figures, explaining: *"we need the witch to stay on this side (of the bridge) – remember the story?",* and asking who the figures are. Levels of educational talk in the museum appear much higher than for the other installations, based on a sample of 10 families. Only four groups did not engage in any educational talk (40%), while two engaged a lot. Four adults talk about the poem and two motivated children to place figures where they should be according to the story (*"There is old Nick – he was IN the churchyard", "put them in the church"*). Others mention the poem's title, its main character, and other figures (*"the witch", "old Nick"*). Often adults narrated during the animation which shows Tam walk past the spooky creatures. An older woman explains: *"that's him coming in… he comes out of the pub, and he had some, some ladies dancing in the church*", she points at a figure, avoiding to mention it is witches in flimsy dresses: *"see there, that's basically in the wrong place"*. Another lady comments*: "See the church he used to walk past - he had to get home, to cross the bridge to get to his house"*.

Moreover, the museum setting provides extra opportunities for educational talk. Spooky Stories is set next to a display with a wooden carving panel of Tam O'Shanter and other exhibits related to the poem. Several adults took a child aside, pointed at the panel and explained the story (fig. 6 top right): *"look, here is the witches trying to catch him, and here they dance"*. Furthermore, the location of the museum invites references. Adults remind children that the poem is set in the same village they are in. A woman with three girls points at the church in the scene and says *"we went there on Sunday to the church"*. Another lady comments the animation *"there's the brig (bridge), remember we were up there, do you remember that?"*.

Scaffolding Children

In the user study of PhotoFit adults always scaffolded children, in particular young ones, read out instructions, repeatedly commented on choices, and were highly involved. But in the museum, only one family out of 18 did some scaffolding, only one adult read out instructions, and levels of emotional support were low. Moreover, in the user study, children frequently looked to adults for approval of choices. This was almost never the case in the museum, where children had to call for adults to look at the face created. User study scaffolding levels varied for the PoetryGame. Half the children did not require any help. For the other half, parents scaffolded. All adults gave extensive emotional support, praised and celebrated high scores (*"well done"*). Of 17 families in the museum, 6 scaffolded lightly (*"get the horse to eat the carrots"*), and 2 strongly (*"OK press, look here, press the button"*). Only five families (30%) motivated and praised (*"your personal best"*). Interaction styles in both settings are more similar for the PoetryGame (albeit with lower rates of parental involvement) than for PhotoFit. Museum scaffolding levels for Spooky Stories appear similar to PoetryGame. Of the sample of 11 families, about half (six) provide scaffolding, read out instructions, and show how to move the transparency or scenery.

DISCUSSION AND FUTURE WORK

Our focus in this paper has been on the differences in behaviors elicited in a semi-realistic (but nevertheless lab-based) user study and in the wild. Even though this was not

Figure 6. Top: Adults reading out quotes and poems in user study. Grandma at Spooky Stories in museum points out "the bridge we were at". Mother explains the story at adjacent carvings. Bottom: PoetryGame. Waiting to have a go. Child being told to leave.

a controlled experiment, we believe we can compare prevalent behaviors between settings. Game play and interaction structure of prototypes were similar to that of final versions.

We found the largest difference for Photofit, with high parental involvement during the user study (educational talk, scaffolding, keeping children on track), and almost none of this occurring in the museum. Moreover, the user study over-predicted the popularity of this installation, with repeated and lengthy gameplay, whereas visitors in the museum rarely repeated play, often did not finish, and pondered little on their choices in the game. The low level of parental engagement differs from how adults normally attempt to enrich children's museum experience [9, 16, 20, 30], indicating that PhotoFit is perceived primarily as a game. Differences were less marked for the PoetryGame, levels of educational talk in the museum being lower then in the user study, but higher then for Photofit. Museum visitors tended to repeat this game at least once and were eager to finish it. Finally, Spooky Stories had the highest levels of educational interaction in the museum.

The differences between lab-based user study and museum can partly be explained by the focused setting of a user study and its demand characteristics. Parents want to be both good participants and parents. But, interestingly, the museum setting did not uniformly affect behavior, some games being almost as popular as in the lab setting and eliciting similar adult-child interactions. We identified a range of influences, such as proximity of other installations affecting activity patterns, physical setup, and floor layout, which influence visibility of installations, comfort of play, and willingness of parents to let children play alone.

There has been a lot of discussion about the benefits of in-the-wild studies [4, 28] and approaches for situated evaluations or quasi-experiments in the wild. But sometimes practical reasons [22] preclude user studies in the wild, such as costs, time constraints, risks incurred by a field study, lack of sufficiently robust prototypes or of access to the site – or, as in our case, formative studies are needed to improve a system before deployment. It is therefore important to know what lab-based user studies can tell us about in-the-wild usage. Our research presents further evidence of the benefits of in-the-wild studies, illustrating how user behavior differs from that in the lab, and adds to the noted caveats of organized field studies [3]. Nonetheless, our user studies were successful in evaluating collaborative play, and enabled us to identify and remedy usability issues. Running user studies with family groups or in party/living-room labs [18] are thus viable user-testing strategies. Yet results from such studies need to be interpreted with caution - often new usability issues emerge in the wild [23]. Our findings illustrate how much researchers and practitioners benefit from first-hand experience in the field by having a sense of what constitutes 'realistic' behavior. Further research is required to extend our methodological repertoire for emulating not just the environmental [12, 25], but also the social context of the use situation, or to devise new ways of running user studies in-situ [4, 25], (and long-term user studies). Moreover, a combination of prototyping and evaluating in the wild [29] could result in new approaches.

In this paper, we have focused on the three touchscreen installations that we originally user-tested. As future work we aim to expand the scope of research questions, in particular, to investigate interactions between strangers. Furthermore, observation in the museum indicates that elderly users engaged more with some of the installations than we would have expected from the user studies. We hypothesize that elderly users were more strongly affected (inhibited) by the lab-study setting then children, who were happy to be allowed to play. Nevertheless, testing with mature users was important to identify usability issues for this age group. We plan another analysis of our data focusing on elderly adults.

CONCLUSION

We have presented a comparison of user interactions with museum game installations in a user study and in the wild, contributing to a reflective discussion of evaluation methods [2, 4, 26]. While enlisting family groups as participants was successful for investigating how well the games work for groups and identifying usability issues, the sessions' by-invitation, focused character affected user responses notably (see [14]). A subsequent observational study in the museum revealed a very diverse picture of overall patterns of use and social interaction, in particular, differences in family interactions and (grand)parental behaviors.

While it is well known that demand characteristics and the context of use influence user behavior, our work investigates what exactly this means in the museum context and for adult-child interaction. Previously, demand characteristics have been discussed mostly with regard to how these influence the attitude to using a system or make people more willing to do a task [3, 6, 33]. Our study highlights effects on the social interactions among users, in this case between children and adults. Our analysis demonstrates the extent of such effects, with measurable differences in whether and how often certain behaviors occurred. Furthermore, we have shown that these effects were not uniform across installations, and how interactions were influenced by factors such as physical placement and installations setup as well as contextualization of the systems. Systematic comparisons of user behavior, such as ours, in different domains and settings, are important to attain a better understanding of the issues that should be kept in mind when interpreting the outcomes of lab-based user studies.

Early user studies and evaluations are a central part of user-centered design. They are indispensible, but often such early prototypes are not ready for a naturalistic user study in the wild. We cannot resolve this tradeoff, but believe that we have contributed to a better understanding of the limitations of organized user studies, allowing us to anticipate and reflect better on how behavior will differ from the wild.

Acknowledgments
We thank the National Trust Scotland, the Kelvingrove Museum, the Robert Burns Birthplace museum, our study participants, and Tobias Fischer for technical assistance.

REFERENCES

1. Allen, S. Designs for Learning: Studying Science Museum Exhibits That Do More Than Entertain", *Science Education Vol. 88(S1),* (2004) S17-S33.
2. Barkhuus, L., Rode, J. From Mice to Men - 24 years of Evaluation in CHI. *alt.chi 2007*
3. Brown, B., Reeves, S., Sherwood, S. Into the wild: Challenges and opportunities for field trial methods. *Proc. of CHI'11*. ACM, NY (2011), 1657-1666
4. Consolvo, S., et al. Conducting In Situ Evaluations for and With Ubiquitous Computing Technologies. *Int. J. of Human-Computer Interaction 22(1),* (2007), 107-122
5. Crabtree, A. Design in the absence of practice: breaching experiments. *Proc. of DIS'04*, ACM (2004), 59-68
6. Draper. S.W. The notion of task in HCI. *Proc. of INTERACT '93 & CHI '93.* ACM, NY (1993), 207-208.
7. Falk, J.H., Dierking, L., Adams, M. Living in a Learning Society: Museums and Free-choice Learning. In Macdonald, S. (ed.) *A Companion to Museum Studies*. Blackwell Publishing (2006)
8. Greenberg S., Buxton, B. 2008. Usability evaluation considered harmful (some of the time). *Proc. of CHI '08*. ACM, NY (2008), 111-120.
9. Gutwill, J.P., Allen, S. *Group Inquiry at Science Museum Exhibits. Getting Visitors to Ask Juicy Questions*. Exploratorium / Left Coast Press (2010)
10. Hagen, P., Robertson, T., Kan, M., Sadler, K. Emerging Research Methods for Understanding Mobile Technology Use. *Proc. of OzCHI'05*. ACM NY (2005), 1-10
11. Hindmarsh, J., Heath, C., vom Lehn, D., Cleverly. J. Creating Assemblies in Public Environments: Social Interaction, Interactive Exhibits and CSCW. *CSCW 14(1):* 1-41 (2005)
12. Hoggan, E., Brewster, S., Johnston, J. Investigating the effectiveness of tactile feedback for mobile touchscreens. *Proc. of CHI '08*. ACM NY (2008), 1573-1582.
13. Hornecker. E. Interactions Around a Contextually Embedded System. *Proc. of TEI'10*. ACM (2010) 169-176
14. Hornecker, E., Nicol, E. Towards the Wild: Evaluating Museum Installations in Semi-Realistic Situations. *Rethinking Technology in Museums Conference 2011.*
15. Hornecker, E. Stifter, M. Learning from Interactive Museum Installations About Interaction Design for Public Settings. *Proc. of OzCHI'06*. ACM (2006), 135-142
16. Humphrey, T. et al. *Fostering Active Prolonged Engagement. The Art of Creating APE Exhibits.* Exploratorium / Left Coast Press (2005)
17. Isbister, K. Enabling Social Play. A Framework for Design and Evaluation. In: Bernhaupt, R. (ed) *Evaluating User Experience in Games.* Springer (2010) 11-22
18. Isbister, K., Schaffer, N. *Game usability: advancing the player experience.* Morgan Kaufmann (2008)
19. Jordan, B., Henderson, A. Interaction Analysis: Foundations and Practice. *J. of Learn. Sc. 4, 1* (1995), 39–103.
20. Kelly, L., Savage, G., Griffin, J., Tonkin, S. *Knowledge Quest: Australian Families Visit Museums.* Australian Museum and National Museum of Australia (2004)
21. Kjeldskov, J., Graham, C. A Review of Mobile HCI Research Methods. *Proc. of MobileHCI'03*. Springer (2003), 317-335
22. Kjeldskov, J., Skov, M., Als, B., Hoegh, R. Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. *Proc. of MobileHCI'04*. Springer (2004), 61-73.
23. Marshall, P., et al. Re-thinking 'multi-user': An in-the-wild study of how groups approach a walk-up-and-use tabletop interface. *Proc. of CHI'11*, ACM: NY (2011) 3033-3042
24. Olsen Jr., D. Evaluating User Interface Systems Research. *Proc ACM UIST'07*. ACM: NY (2007) 251-258.
25. Oulasvirta, A. Field experiments in HCI: Promises and challenges. In P. Saariluoma, H. Isomaki (Eds.), *Future Interaction Design II*. Springer 2009
26. Pinelle, D., Gutwin, C. A Review of Groupware Evaluations. *Proc 9th IEEE Int'l Workshop on Enabling Technologies WET-ICE'00*. (2000). 86-91.
27. Plowman L., Stephen, C. McPake, J. *Growing Up with Technology: Young children learning in a digital world*. London: Routledge 2010
28. Rogers, Y., et al. Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. *Proc. of Ubicomp'07*, ACM (2007)
29. Rogers, Y. Interaction Design Gone Wild: Striving for Wild Theory. *interactions 4(8),* (2011). 58-62
30. Sanford, C., Knutson, K., Crowley, K. "We Always Spend Time Together on Sundays": How Grandparents and Their Grandchildren Think About and Use Informal Learning Spaces. *Visitor Studies 10:2,* (2007) 136-151
31. Smith, R., Iversen, O.S. When the Museum Goes Native. *interactions 5(8),* (2011), 15-19
32. Snibbe, S., Raffle, H. Social Immersive Media: Pursuing best practices for multi-user interactive camera/ projector exhibits. *Proc. of CHI'09*, ACM (2009). 1447-1456
33. Snyder, C. *Paper Prototyping*. Morgan Kaufmann 2003
34. vom Lehn, D., Hindmarsh, J., Luff, P., Heath, C. Engaging Constable: Revealing art with new technology. *Proc. of CHI'07*. ACM (2007). 1485-1494