Understanding Public Evaluation: Quantifying Experimenter Intervention

Julie R. Williamson and John Williamson
University of Glasgow
Glasgow, Scotland, United Kingdom
{Julie.Williamson, JohnH.Williamson}@glasgow.ac.uk

Figure 1. Public evaluations are difficult and messy. The real world experience captured in public spaces is both integral to authentic experience and challenging for researchers. Tedious hours can be spent waiting for potential users to approach a public display (left), the activities of others in public space can be disruptive (centre), and serious hardware failures can mean failed data capture (right).

ABSTRACT
Public evaluations are popular because some research questions can only be answered by turning "to the wild." Different approaches place experimenters in different roles during deployment, which has implications for the kinds of data that can be collected and the potential bias introduced by the experimenter. This paper expands our understanding of how experimenter roles impact public evaluations and provides an empirical basis to consider different evaluation approaches. We completed an evaluation of a playful gesture-controlled display – not to understand interaction at the display but to compare different evaluation approaches. The conditions placed the experimenter in three roles, steward observer, overt observer, and covert observer, to measure the effect of experimenter presence and analyse the strengths and weaknesses of each approach.

Author Keywords
Public Evaluation; Public Displays; In the Wild Methods.

ACM Classification Keywords
H.5.2. User Interfaces: Evaluation/Methodology.

INTRODUCTION
In meteorology, the butterfly effect refers to small changes - metaphorically the fluttering of a butterfly's wings - having a large knock-on effect on the weather. For interactive public technology, the equivalent might be the very literal distraction of a passing butterfly suddenly drawing away users' attention from an installation. The delicate and unpredictable relationship between public technology, the environment in which it is situated, and potential users calls for careful "in the wild" study to authentically capture experiences. Such evaluations are difficult to control, as shown in Figure 1, and experimenter intervention is often required. Experimenter interventions can also reveal valuable insights into behaviour and motivation, but represent a very large butterfly indeed.

Some research questions can only be answered through real world evaluation [20]. Results found in lab studies alone may not be a good predictor of usage in real world contexts [14]. Investigations of *user experience*, as opposed to usability, need "in the wild" methods to understand complex social and physical contexts, explore how experiences develop over time, and uncover new and unexpected uses [15]. There are diverse approaches to such evaluations, and even the meaning of "in the wild" is not widely agreed [18]. It is often used interchangeably with field trials [1,4], deployments [1], and intervention-based studies [34]. In this paper, we restrict our scope to "in the wild" studies of stationary technology in public and semi-public places, for example public displays and interactive installations. Crucially, this involves *intervening* (through the installation of technology) within a *real world place* (not a lab simulation). We refer to studies of these interventions as *public evaluations* throughout this paper.

Popular approaches to *public evaluations* include controlled on-the-street studies [6,25], observational studies [23,33], and steward observation [29]. However, there is limited work reflecting on the bias implications of these

CHI 2017, May 06-11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4655-9/17/05…$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025598

approaches. Many investigators decide on approach based on personal experience, preference, and perceptions of reviewer acceptability. Foundational work reflecting on "in the wild" methods [for example 3,11,20] provides valuable insights, but the field lacks systematic empirical work that investigates the implications of *experimenter roles* on data collection. Debates about approach focus on how much the experimenter should intervene during evaluation, ranging from overtly creating a controlled experience (for example [6]) to completely unstewarded experience (for example [35]). The key question for public evaluation is: how does experimenter intervention distort results?

This is a foundational problem throughout science, but *public evaluations* are particularly fragile with respect to intervention bias.

Demand characteristics In public evaluations users' behaviours are significantly influenced by *demand characteristics* [4], a bias introduced when participants act to satisfy perceived desires of the experimenter. A lonely looking researcher in front of a display can attract pity interactions that do not represent genuine desires to interact.

Unfamiliarity Public evaluations often introduce new concepts that lack existing practice [8] where investigator presence and instruction will significantly influence how participants react. The pose and subtle reactions of an experimenter might be the only cues a potential user can quickly grasp at an unfamiliar, experimental art installation.

Shyness Finally, public evaluations often explore voluntary interactions, which users may quickly abandon in the face of difficulty, discomfort, or external pressure. A passer-by's fleeting curiosity about a new street installation might easily be diverted by the gaze of a watching observer.

Anecdotal evidence indicates presence of an experimenter introduces bias, but the magnitude and attributes of these effects have not been quantified for public evaluations. This is exacerbated by pressure to present public evaluation studies as predictable, replicable, and comparable [4], which diminishes or ignores sources of bias.

This paper provides an empirical basis to consider the effect of experimenter roles in different evaluation approaches. To do this, we evaluated a playful gesture-controlled display, controlling not the interaction techniques but the experimenter roles. The conditions had experimenters take on three roles: steward observer; overt observer; and covert observer. We combined this with high-density quantitative measurements to precisely quantify the impact of experimenter presence. For this first time, this captures how experimenter presence introduces bias and documents anomalies arising from observation.

This paper:

- Identifies the core methodological challenges of public evaluations
- Details an empirical evaluation of the impact of three common public evaluation approaches in practice
- Shows that experimenter role has a dramatic and clearly measurable effect on how people use public installations
- Lays out a systematic experimental framework for investigating observation roles in public evaluation.

CHALLENGES FOR PUBLIC EVALUATIONS

Although there are clear benefits to evaluating "in the wild", real world deployments are hard [4]. Public evaluations are logistically complex and resource intensive, especially compared to traditional lab evaluations [17]. Researchers need to work with project partners and external organisations to negotiate site access, install equipment, and agree on issues such as security, visual identity, electricity, and ethics. Researchers face significant engineering challenges to create quality prototypes that users will take seriously, and robust enough prototypes that users won't break them. Wild users respond in unusual and expected ways, resulting in obscure errors, data collection failures, and even deliberate vandalism. This is compounded by the fact the studies often run over long periods of time – days or weeks rather than hours or minutes.

Solving these practical issues is essential. But, more importantly, the *experimental design* must be capable of capturing aspects of authentic behaviour. This needs to be solved before trials start: there are often no second chances. The event that supported the evaluation may have passed, the deployment site may no longer be available, and resources may simply be spent, or word may spread that an installation is broken or buggy. Despite this, public evaluations are "worth the hassle" [27] because of the rich and otherwise inaccessible results they offer. However, there is little empirical work that informs researchers on the impact of different approaches. There is a lack of theory to guide "in the wild" evaluations [26], and evaluating new interactions can be difficult when users have no idea what to expect [8,9]. The biases introduced can be severe and unpredictable [4], making it difficult to produce robust and replicable results.

We organise the *methodological* challenges in public evaluation into three themes; *experience control*, *authenticity,* and *ethics*.

Experience Control

Public evaluations require researchers to give up direct control of interactions. It is this shift of control from experimenter to participant during "in the wild studies" [26] that makes experimental design so much more challenging than traditional lab studies. For example, participants cannot be treated as equivalent and interchangeable units, conditions may be unbalanced, and difficult to identify confounding factors pollute results.

Experimental design means balancing control with ecological validity [16]. Increased intervention increases control but also increases bias. For example, *demand*

characteristics [4] will be amplified where participants have more contact with the experimenter. The verbal framing of on-the-street studies like [6,25] can significantly influence results [4]. *Anomalies*, or unplanned interruptions during public evaluations, are an essential part of the experience of public technologies but introduce unpredictability into data collection. For example, in our study, we encountered a truck that blocked views from recording equipment, a procession that drew attention from the display, friends of the experimenter approaching and engaging in poorly-timed conversations, and the distracting activity of nearby animals.

When discussing experimental control, our evaluation addresses the following questions:

- Does asserting control during an evaluation change users' behaviours?
- Can desired behaviours be elicited without experimenter intervention?

Authenticity

We do "in the wild" evaluation to collect ecologically valid data [5,7]. But defining what "valid" or "authentic" data constitutes in public evaluations is hard and the research community lacks a shared ideal of validity [1]. There are strong but conflicting arguments for covert methods [34], mixed methods [33], and participant observations [15,29].

Questions about authenticity often focus on qualitative versus quantitative data and there are strong opinions on both sides. While quantitative comparisons may not capture important aspects of "in the wild" experiences [27], qualitative results are hard to generalise and may be difficult to reproduce. The push for design recommendations and other "generalizable" results from qualitative work also over simplifies such methods in a harmful way, failing to capture the strengths of in depth qualitative inquiry [10]. Rejecting pressure to make "in the wild" studies objective and reproducible [4] with neatly packaged results [10] leads to a mixed methods approach that embraces the variability of public evaluations.

Researchers have a wide spectrum of tools to collect data from experiments. These include usage logs, video and audio recordings, participant observation, on-the-street interviews, and overt or covert observation. There are a range of techniques for analysing these data streams, for example pedestrian tracking [36], f-formations [19], and the audience funnel framework [22].

Understanding non-use and avoidance is essential in situated technologies [2], and requires data collection methods that support this. This helps to mitigate the bias of *lead participants* [4], participants whom actively help the trial and/or perform exemplary interactions, by putting the rarity of their interactions into context.

This paper explores the authenticity of data when results differ for different evaluation styles. Our discussion revolves around the following questions:

- What should be the considered the "ground truth" for real world interactions?
- Are results biased when analysis favours interacting users, ignoring or downplaying the role of non-interacting users?

Ethics

There are significant practical and theoretical ethical challenges in public evaluations [32,34]. The core ethical issue is that capturing authentic experience may demand that participants not be aware they are involved in an evaluation [32][34].

Public evaluations often blur the boundaries between public events and research [12], placing investigators in a complex ethical situation. Practical challenges such as gaining appropriate ethical approval, choosing whether to gather explicit informed consent from participants, and ensuring participant safety have been discussed thoroughly in previous work [12,32,34]. Methodological challenges around recruitment, motivation, and intervention are harder to address. Reilly et al. [24] discuss the ethics of using public events to access participants who may otherwise be uninterested in participating in research evaluations. Waern [32] discusses the importance of recognizing that public evaluations are interventions that inevitably alter public spaces that people are actively using, perhaps for the worse.

We analyse the specific ethical issues that arise when experimenters take on different roles during evaluation, addressing the following questions:

- What are the ethical issues of intervening with face-to-face encounters in public spaces?
- To what extent does unwitting participation undermine a passer-by's right to refuse participation?

PUBLIC EVALUATION APPROACHES

Approaches to public evaluations are incredibly varied, and of the key decisions is the role of the investigator. Investigator roles fall into three main categories; *steward observer*, *overt observer*, and *covert observer*. This has implications for the types of data that can be collected (covert observation precludes interviews, for example) and the behavioural bias induced. Figure 2, organises the prominent display-based and tangible interface examples from the literature according to the type of data collected (qualitative and quantitative) and the level of experimental control (controlled and uncontrolled).

Steward Observer

As a *steward observer*, the investigator takes an active role in managing, curating, and controlling the experience of the participant. Unlike ethnographic *participant observation,* steward observers take on a different role from users,

actively presenting themselves as stewards, owners, or performers.

Researchers took an immersive role in *humanaquarium* [17], an interactive performance where the audience could participate through a touch sensitive enclosure around the performers. Touch input during the performance changed the audio output of the performers' instruments and the visual projections behind the performers. The *humanaquarium* researchers used their experiences from performing to inform the design of the installation.

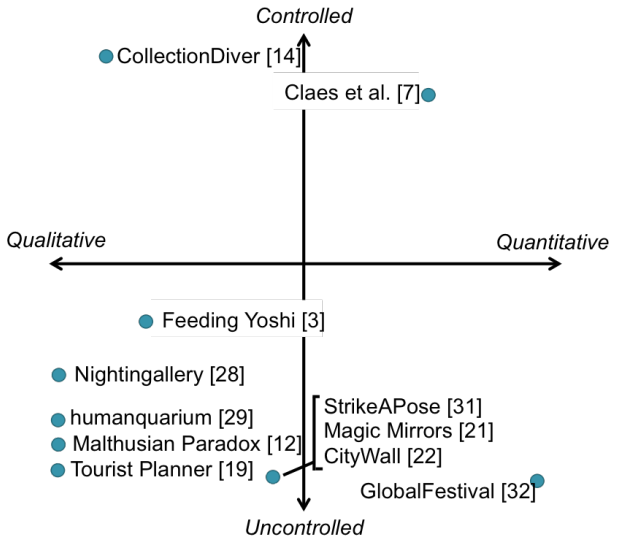

Figure 2. The evaluation studies discussed in this paper are organised based on experimental control (controlled to uncontrolled) and the primary data collected (qualitative to quantitative).

Nightingallery [30] positioned researchers as circus barkers, attracting and guiding users to interact with an animatronic bird. In both *humanaquariam* and *Nightingallery* the researchers provided guidance in an open-ended interaction.

In contrast, Claes et al. [6] used steward observers to recruit passers-by in a public space to complete a series of tasks with the public tangible display. Hinrichs et al. [13] completed an *in situ* study of a tangible library search interface called *CollectionDiver*, but rather than passers-by, explicitly recruited participants. Participants completed a series of tasks with the interface and were then interviewed.

Steward observation allows researchers to produce in depth, qualitative results, where researchers can probe specific questions and elicit specific behaviours.

Overt Observer

An *overt observer* observes without actively intervening during interaction, although survey or interview data may be gathered after interaction. Overt observation is often used to add qualitative results to interaction logs and video analysis, but the impact of the overt observation on this data has not previously been evaluated. For example, passers-by may be deterred by overt observers or feel uncomfortable when approached after interaction.

In *City Wall* researchers observed users at a multi-touch display [23] and used on-the-street interviews to complement observation notes and video analysis. A similar approach was used to evaluate the *StrikeAPose* gesture controlled display [33] and *Magic Mirrors* [22]. In each case, on-the-street interviews were gathered by approaching users after they had interacted with the display.

However, the selection bias effects (diverting shy users) of overt observation on interaction is not currently known. The ethics of approaching users in this way is also unclear, where they may regret interactions when they realise they are involved in a study.

Covert Observer

As a *covert observer*, the investigator does not maintain any visible presence or intervene in any way during evaluation. This approach relies on quantitative and qualitative analysis based on video or audio recordings and interaction logs.

The *GlobalFestival* [35] information display was evaluated completely through video analysis, combining automatic pedestrian tracking [36] with detailed manual analysis of interactive segments. *The Tourist Planner* [21] table top display was evaluated using a combination of overt observation (first phase of study) and covert observation (second phase of study) through video data. Covert observation is similar to evaluations of pervasive mobile technologies, for example *The Malthusian Paradox* [11] and *Feeding Yoshi* [3]. These both recruited participants in advance and asked them to use pervasive applications in everyday life without the experimenter present. However, pervasive games grant different agency to users than public evaluations. In pervasive games, users are explicitly recruited and knowingly participate, choosing when and where they want to interact.

Although covert observation removes the possibility of gathering on-the-street responses from users, advocates argue that this represents the most ecologically valid experience [34]. Because public displays have such a delicate balance between interaction and non-interaction, overt and stewarded observation may seriously distort the self-selection of interacting users.

UNDERSTANDING PUBLIC EVALUATION

We set out to quantitatively capture the effects of experiment roles on the interaction/non-interaction bias and distorting effects on the interaction experience. We developed *Silly Hats Only*, a playful gesture controlled display. We performed multiple evaluations of this display, with the experimenter role as the controlled independent variable.

Silly Hats Only

Silly Hats Only is a playful installation where users can see their silhouettes as captured by a depth camera, as shown in

Figure 3. Physically, it consists of a large freestanding display in a busy public walkway. When they make a teapot gesture (putting your hand on your hip), similar to the interactions used in *StrikeAPose* [33], a silly hat is drawn upon their head while they maintain the teapot pose. When users approach the display, their silhouette appears alongside an animation demonstrating the teapot gesture.

Figure 3. When a user performs the teapot gesture, their silhouette will get a silly hat. An animation in the bottom left of the screen demonstrates the gesture.

The goal was *not* to analyse the interaction, but to create a study scenario that captured the essence of public evaluation. We used this to evaluate the impact of the experiment role directly.

Our core question is: **How do experimenter roles influence the data collected during a public evaluation?**

To answer this, we contextualised this question in terms measurable variables relevant to our installation:

Q1: What are users' preferred interaction distances when playing with a gesture-enabled display?

Q2: How do higher rates of pedestrian traffic on the walkway influence interaction distance?

Experimental Design

Three conditions placed the experimenter in different roles; steward observer, overt observer, and covert observer.

Steward observer

The steward observer condition was based on the approach developed by Claes et al. [6]. The steward observer stood close to the display and recruited participants from passers-by. Participants were recruited if they 1) approached the experimenter or 2) walked close enough to be captured by the depth camera and showed an interest in the display. Each participant was then asked to perform a series of tasks and answer questions about their experiences.

In this condition, the following data was captured:

- Structured Usage Data (task timelines)
- Quantitative Responses (interaction logs)
- Qualitative Responses (interview questions)
- Overhead Video (tracking pedestrian motion)

Overt Observer

The overt observer condition had the experimenter seated roughly five meters from the display (within the overhead video frame). The experimenter kept an open notebook visible and sat oriented towards the display. The experimenter did not intervene, but would answer questions if approached. Although many overt observer studies include follow-up questions with passers-by who have engaged, we chose to exclude this to minimise disruption.

This condition gathered the following data:

- Quantitative Responses (interaction logs)
- Observation Notes (in person)
- Overhead Video (pedestrian motion tracking)

Covert Observer

The covert observer condition had no visible experimenter. The experimenter gathered observation notes from a live video stream from indoors.

This condition gathered the following data:

- Quantitative Responses (interaction logs)
- Observation Notes (from video, same perspective as pedestrian tracking)
- Overhead Video (pedestrian motion tracking)

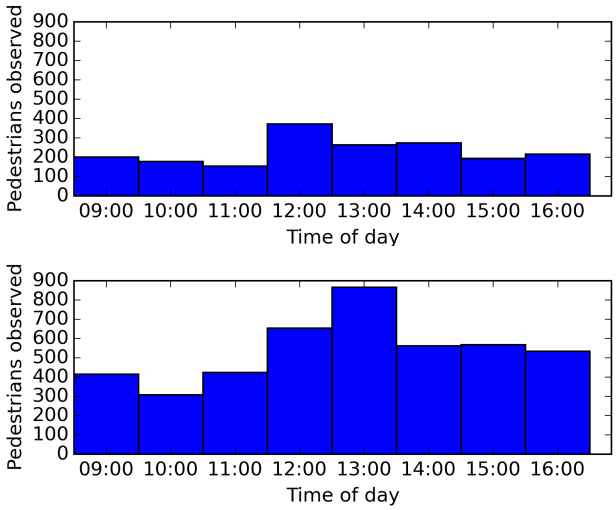

Figure 4. Passers-by per hour on low-traffic day (top) and high traffic day (bottom) during a special event. Units are minutes from 9am.

Hardware and Setting

The evaluation used a 42" high definition screen in a waterproof grey "monolith" enclosure, as shown in Figure 3. Depth images were captured using a Kinect 2.0 device[1]

[1] Kinect 2.0

https://developer.microsoft.com/en-us/windows/kinect/develop

placed below the display and processed using the OpenNI libraries[2]. The experimental software logged all interaction events and position data for all user silhouettes captured by the depth camera. Overhead video, from a camera positioned 15m above and 15m behind the installation, was captured for qualitative and quantitative analysis. The display was positioned in an outdoor semi-pedestrianised walkway that had regular foot traffic throughout the day, as shown in Table 1.

Table 1. Total numbers for passers-by and users observed during each condition. Users are counted as all silhouettes captured by the depth camera. Users are broken down into more refined categories in Figure 10.

	Total Hours	Total Passers-by	Total Users
Baseline	20	6810	N/A
Steward Observer	4	1634	401 (25%)
Overt Observer	4	1769	378 (21%)
Covert Observer	4	2000	570 (29%)

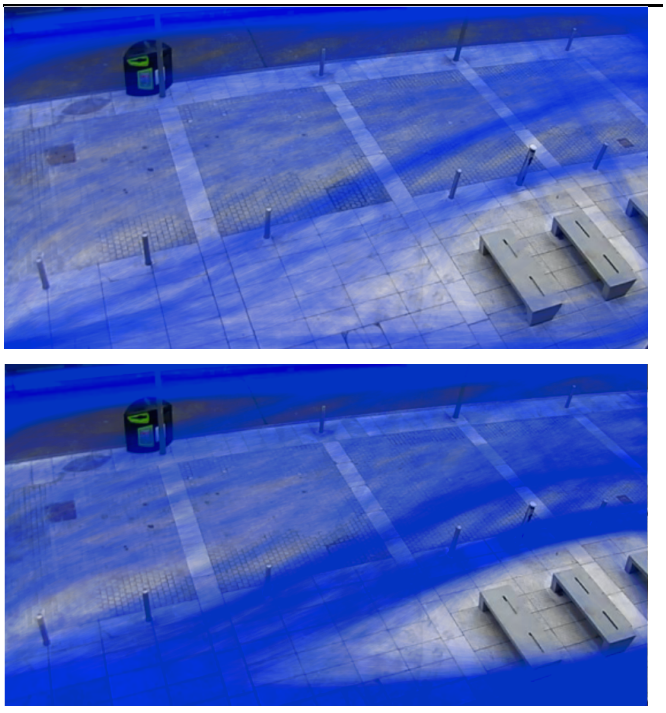

Figure 5. Baseline pedestrian motion in the space used for evaluation. Each blue line represents the motion of one pedestrian through the space [extracted automatically from video]. Top: Low traffic baseline, N=1993, Bottom: High traffic baseline, N=4817. Pedestrian routes are similar in both high and low traffic days.

[2] OpenNI - http://structure.io/openni

RESULTS

Data was collected in two separate weeks to capture "high" and "low" pedestrian traffic levels, as shown in Figure 4. Six hours of interaction data (two one-hour blocks per condition) and ten hours of baseline data were collected during a "low traffic" period (typical footfall) and six hours of interaction and ten hours of baseline data during a "high traffic" (a special event). The order of the condition (experimenter role) blocks was randomised, with the display installed during daytime hours between 10:00 and 18:00.

These results are based on a total twelve hours of interaction data (four one-hour blocks for each of the three conditions) and twenty hours of baseline pedestrian motion data (two ten hour blocks) gathered in the walkway. The baseline data was captured without any installation present.

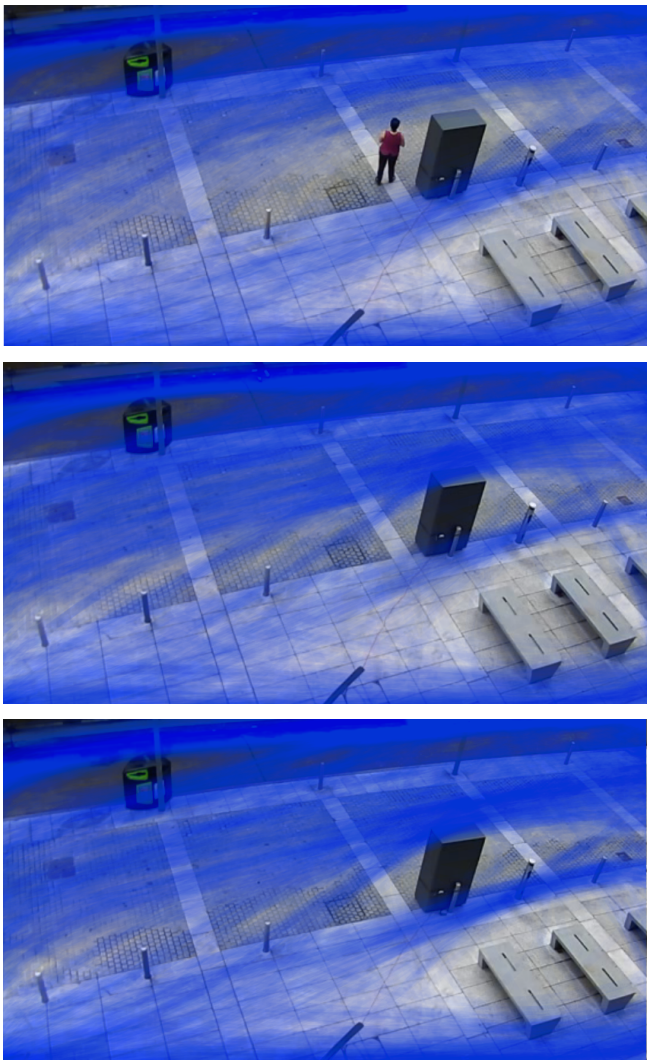

Figure 6. Pedestrian traffic for three evaluation conditions. Each blue line represents the motion of one pedestrian through the space [extracted automatically from video]. Top: Steward Observer, N= 1634, Middle: Overt Observer, N=1769, Bottom: Covert Observer, N=2000.

Figure 4 shows the footfall per hour during the baseline datasets and Figure 6 shows visualisations of the pedestrian traffic on the baseline days. Although footfall is roughly doubled during the high traffic period, patterns of routes between low and high traffic days are similar. Table 1 gives an overview of the raw data.

Pedestrian Traffic

The pedestrian data was generated using open source pedestrian tracking software [36]. Figure 6 visualises how the display disrupts the pedestrian traffic in the walkway, with many passers-by walking behind the display or giving a wider berth in front. In the baseline data in Figure 5, a common route runs through the walkway near the third bollard on the left, but this is dramatically reduced when the installation is present as passers-by swerve to avoid the display.

The area directly in front of the display where interaction can occur is still a busy thoroughfare. At peak times, interacting users would likely be "in the way" of others moving through the walkway. Many passers-by walked close enough to be captured by the depth camera but did not stop to interact and were unlikely to be aware they had even passed through the interactive zone. We further explore different "interaction styles" based on interaction duration and engagement with the activation gesture in the "Interaction Profiles" section of this paper.

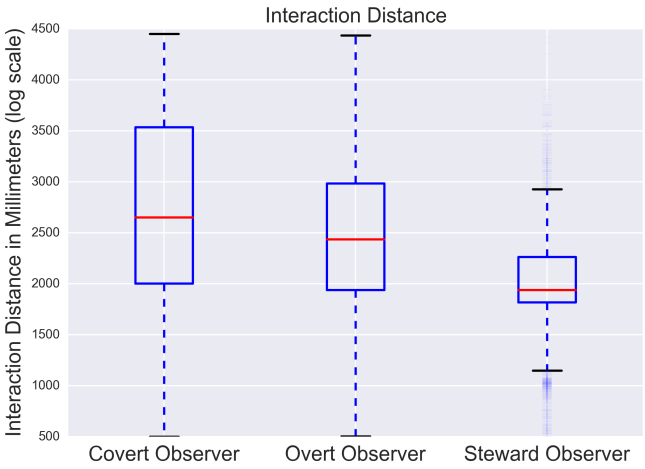

Figure 7: Average standing distance from the display for each condition in millimetres.

Pedestrian traffic was most disrupted in the steward observer condition, where passers-by clearly avoided the entire area around the experimenter. This disruption has ethical implications, as the intervention potentially made passers-by uncomfortable in the walkway. Such disruption also influenced non-interaction and participant self selection, potentially changing the kind of participants that would be attracted to the display.

Answering Q1 and Q2

For Q1, each condition resulted in different standing distances. Users' average standing distance was 2642 mm for covert, 2360 mm for overt, and 1966 mm for steward, as shown in Figure 7. Using a two-tailed t-test for this normally distributed data and Cohen's *d* for effect size, pairwise comparisons show that each condition is significantly different.

- covert-overt, $p<0.001$ and $d=0.3$
- covert-steward, $p<0.001$ and $d=0.8$
- overt-steward, $p<0.001$ and $d=0.5$

The different between covert and overt conditions has a medium effect size (>0.2), indicating these differences would be difficult to observe without statistical analysis. However, the differences between covert-steward and overt-steward conditions have a large effect size (>0.4), indicating these differences could be easily observed with the naked eye. The average difference spans over 650mm between covert and steward conditions, which covers a substantial different in a busy walkway.

These results demonstrate a significantly skewed result. For example, if the steward observer condition alone was used to determine how closely a gesture controlled display could safely be placed near a bicycle path, users would likely stand in the bicycle path during interaction based on the typical overt or covert interaction distances. In this case, the steward observer fails to provide results that reflect users' realistic behaviour.

Users in the steward condition were also much more uniform in their standing positions and interacted in a much smaller range of areas. The presence of the experimenter stabilised interaction and removed randomness in the data (for example from passers-by walking around interacting users during overt and covert observations). This stabilising effect removes some of the realism of that data and fails to capture the noise and bustle of the walkway.

For Q2, qualitative data indicated that users were worried about "blocking the walkway" which explains their closer standing distance to the display. This, however, did not reflect what people did in practice. Observation notes gathered during the covert and overt conditions indicate that user did not seem to worry about blocking others in the walkway, often positioning themselves in the centre of the thoroughfare. Figure 8 shows some examples of how people positioned themselves while interacting. Figure 9, top, shows a large group blocking a large section of the walkway while playing with the display. Figure 9, middle, shows a user making shadow puppets before calling over another user to show him the display. Figure 9 bottom, shows one user who repeatedly looked back at friends seated behind her while she interacted with the display.

Interview Data

One advantage to steward observation is the opportunity to gather qualitative feedback during on-the-street interviews. Although some users' feedback could have been derived from qualitative analysis of the overt or covert video logs,

some insights could clearly only have been gathered through direct questioning.

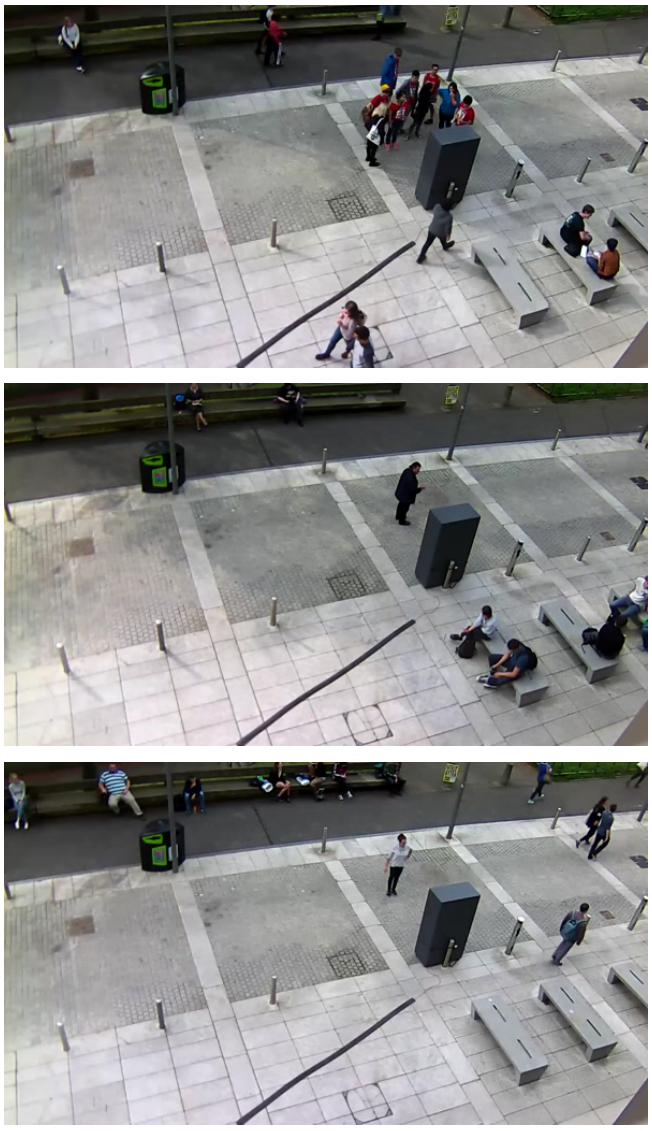

Figure 8. Users often interacted with the display in the centre of the walkway. Top: A large group blocks the majority of the walkway. Middle: Shadow puppets amuse a passer-by. Bottom: A user performs the gesture with an audience watching.

Participants discussed elements of the display's functionality, for example that it was possible to interact while encumbered and that multiple users could interact simultaneously. Participants also frequently mentioned that they would stand as close to the display as possible to prevent blocking the walkway (although as we have seen, they did not do this in practice). One passer-by mentioned that she was returning to the display after seeing it previously to show her daughter because it reminded her of something from her home city. Understanding how her previous experiences led her to interact with *Silly Hats Only* in multiple sessions could not have been understood from video analysis alone. Overall, participants stated that although the interactions were silly, they had no reservations about using the display in the crowded walkway. These themes are comparable to results from similar displays, such as *StrikeAPose* [33].

There were some practical issues with running the steward observer condition. This was the most laborious of the three conditions, requiring the experimenter to actively recruit participants and remain energetic for long periods of time while standing at the display. Three passers-by declined to participate after speaking to the experimenter. Of those who did agree to participate, several participants expressed some discomfort during the evaluation. For example, one participant said that "I don't know where we are going with this. I don't know what is the purpose of this." Recruiting people on-the-street has the potential to make user uncomfortable, either by refusing to participate or by participating when they really would rather not.

Interaction Profiles

Users of *Silly Hats Only* can be broken down into different "profiles" of use based on interaction duration and engagement with the activation gesture. Inspired by the Audience Funnel Framework [22] and proxemic interaction [31], we divided all users captured by the Kinect into four categories: *unaware, accidental, curious*, and *engaged*. Our categories are defined such that they can be automatically based on duration of visibility and number of interactions. The Audience Funnel Framework requires data about whether a user glances at the display, or returns after previous interaction. This could be possible in the near future, for example using more advanced skeleton tracking [28], but is not currently reliable using the OpenNI libraries that *Silly Hats Only* was dependent on.

Unaware users were visible for less than five seconds and did not perform the activation gesture. This represents users walking in front of the display without slowing. *Accidental* users were visible for less than five seconds but performed the activation gesture at least once. These users did not stop or slow to view the display and so the activation was likely accidental. *Curious* users were visible for five seconds or more but did not perform the activation gesture. Such users may have slowed to view the display or stopped completely, but failed to understand or notice the activation gesture. Finally, *engaged* users spent five seconds for more at the display and performed the activation gesture at least once.

Figure 9 shows users for each condition grouped in these four categories. By looking at the number of passers-by that convert to active users, the *conversion rate*, we can see significant differences. Using the N-1 Chi-Square test to compare *conversion rates*, there is a significant difference between the conversions rates for all three conditions.

- Overt-covert: $p<0.001$, 100% chance Covert has a higher conversion rate

- Steward-overt: p<0.02, 98% chance Steward has a higher conversion rate
- Steward-covert: p<0.007, 99% chance Covert has a higher conversion rate

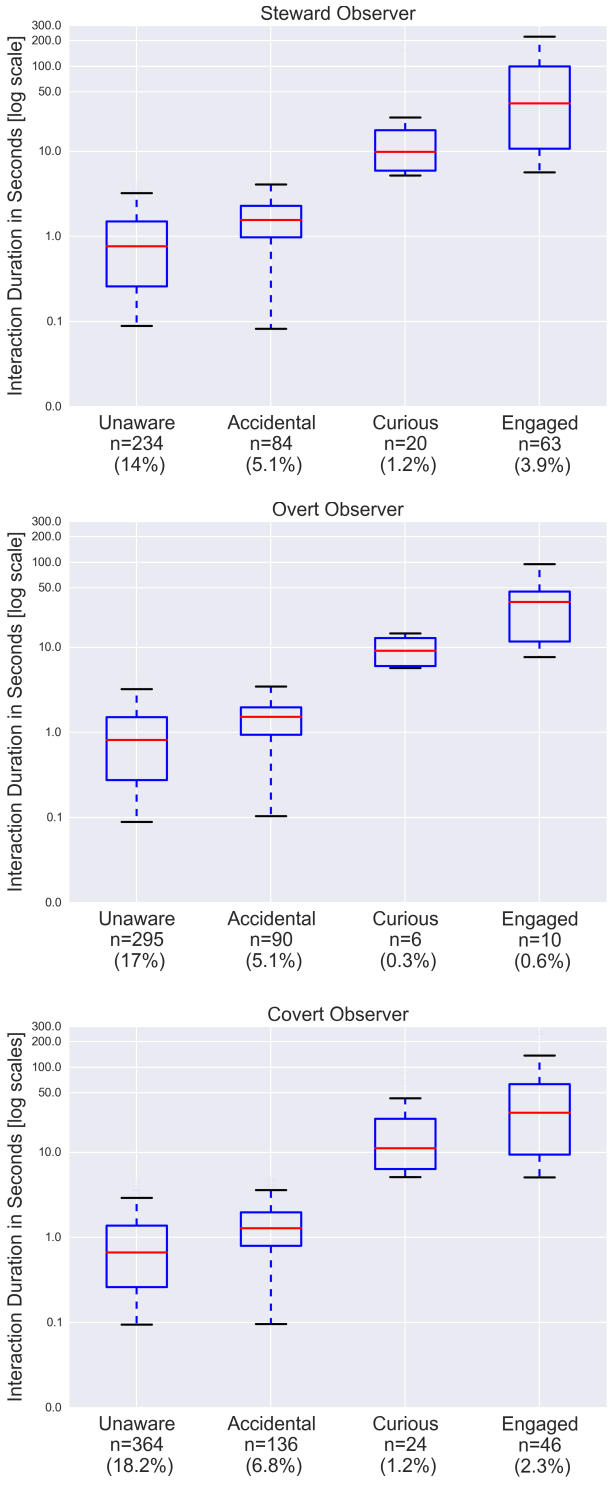

Figure 9. All users have been organised into four categories based on duration of visibility and number of interactions. Percentages show proportion of users out of total observed passers-by.

The most interesting results contrast the behaviour of *engaged* users between the overt and covert observer. The presence of the overt observer significantly influenced *conversation rate,* or the proportion of users that actively engaged with the display. In the Covert condition, 3.5% of passers-by engaged with or were curious of the display. In the Overt condition, this dropped dramatically to 0.9%. Overt observation also disturbed the actions of *curious* users. Users were less likely to explore the display for more than ten seconds if they could not discover how to interact, when they felt they were being observed.

DISCUSSION

Perhaps the most important result is the significantly lowered conversion rates (the proportion of all passers-by that end up spending five seconds or more at the display) when an overt observer is present. In this case, conversion rates dropped to below 1%, from 3.5% in the covert observer condition.

This raises the key question of which users are being driven away? Are they drawn from the same population as users who interacted, or does the observation systematically exclude certain kinds of users? In our experiment, overt observers were unable to effectively "camouflage" themselves, especially during the low traffic conditions when they especially stood out. Future investigation might consider these questions across a finer-grained spectrum of camouflage, with levels between fully overt and covert.

In the steward observer condition, we observed strong "stabilising" effects, leading to much more uniform interactions than when users were left to their own devices. For our case, these stabilising effects were severe enough that the experiment could no longer authentically capture real world usage. Although direct responses captured interesting anecdotes, it arguably did not add significantly to the results. Other experiments, which involve more nuanced interactions or depend more heavily on personal experience, may derive more value from this type of qualitative data.

Our approach lays the foundation for a systematic framework for evaluation observer roles and evaluation methods in public spaces by measuring pedestrian traffic. These results provide an empirical baseline that captures the most common approaches, but further conditions could expand this framework. For example, our study did not include a non-interactive display condition. Additionally, our system was relatively simplistic in its interaction and capabilities. Further research exploring non-interactive displays, non-functioning displays, and more complex displays would significantly expand our understanding of observer effects in a wider range of scenarios.

The significance and magnitude of the observer effects seen even in this simple and playful installation suggest that observation must be very carefully considered in public evaluation. However, in some cases unstewarded

installation may not be practical or meaningful. For example, researchers may be integrated within the installation itself like in *humanaquarium* [29]. In this case, we would be interested in understanding if bias is introduced depending upon the experimenter's mindset; whether they view the installation as evaluation or a performance.

Public evaluations involve complicated and subtle ethical questions, which come to the forefront when there is direct contact between experimenter and (possibly unwitting) participant. In our stewarded condition, the refusal of consent by passers-by and potential discomfort of apparently consenting participants raises ethical questions. Does recruiting people from the street create negative experiences of the public space for some passers-by? Is this counterbalanced by the potential positive effect of an improved installation? Do researchers have a heavier responsibility in safeguarding the public from discomfort than other users of the space (street performers, work men, protesters)? The literature does not address users' reactions to being approached after interaction, but our steward condition suggests that this could potentially be a negative interaction, where an otherwise positive experience may be tinged with regret upon discovering they are part of an evaluation. How can that regret be addressed? Does mere deletion of an individual users' data really address the negative experience of a realisation of unwitting participation?

CONCLUSION

We completed an evaluation of a gesture controlled public display to quantify how experimenter roles distorted results. We used a playful gesture controlled display as a test-bed for different experimenter roles. Our evaluation cast the experimenter as a covert observer, overt observer, and steward observer. We treated the covert observation condition as the ground truth -- the closest approximation of an authentic experience -- capturing users' undisturbed behaviours and willing interactions. The results demonstrate that overt observation significantly reduced interaction rates and discouraged users from exploring the interaction if they initially failed to activate the display. While the active recruitment in the steward observer case maintained interaction rates, it influenced pedestrian movement and significantly altered user behaviour, resulting in artificially uniform interactions at an unnaturally close distance to the display. We propose that systematic control of experimenter roles in public evaluations, and the use of high-density, high-quality measurements like pedestrian tracking are essential in quantifying the observer effect in the fragile and unstable domain of public evaluations. This protocol gives qualitative researchers a way to bracket the authenticity of their results with quantitative, replicable metrics.

REFERENCES

1. Florian Alt, Stefan Schneegass, Albrecht Schmidt, Jörg Müller, and Nemanja Memarovic. 2012. How to evaluate public displays. *2012 International Symposium on Pervasive Displays (PerDis'12)*: #17.
2. Eric P. S. Baumer, Jenna Burrell, Morgan G. Ames, Jed R. Brubaker, and Paul Dourish. 2015. On the importance and implications of studying technology non-use. *Interactions* 22, 2: 52–56.
3. Marek Bell, Matthew Chalmers, Louise Barkhuus, Malcolm Hall, Scott Sherwood, Paul Tennent, Barry Brown, Duncan Rowland, and Steve Benford. 2006. Interweaving Mobile Games With Everyday Life. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06)*: 417–426.
4. Barry Brown, Stuart Reeves, and Scott Sherwood. 2011. Into the wild: Challenges and opportunities for field trial methods. *SIGCHI Conference on Human Factors in Computing Systems (CHI'11)*, May: 1657–1666.
5. Scott Carter, Jennifer Mankoff, Scott R. Klemmer, and Tara Matthews. 2008. Exiting the Cleanroom: On Ecological Validity and Ubiquitous Computing. *Human–Computer Interaction* 23, 1: 47–99.
6. Sandy Claes, Niels Wouters, Karin Slegers, and Andrew Vande Moere. 2015. Controlling In-the-Wild Evaluation Studies of Public Displays. In *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems*, 81–84.
7. A Crabtree and A Chamberlain. 2013. Introduction to the special issue of "The Turn to The Wild." *ACM Transactions on ...* 20, 3: 0–3.
8. Andy Crabtree. 2004. Design in the absence of practice: breaching experiments. *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*: 59–68.
9. Andy Crabtree. 2004. Taking technomethodology seriously: hybrid change in the ethnomethodology–design relationship. *European Journal of Information Systems* 13, October 2003: 195–209.
10. Paul Dourish. 2006. Implications for Design. *SIGCHI conference on Human Factors in computing systems.\n*: 541–550.
11. Elizabeth Evans, Martin Flintham, and Sarah Martindale. 2014. The Malthusian Paradox: performance in an alternate reality game. *Personal and Ubiquitous Computing* 18, 7: 1567–1582.
12. Dustin Freeman, Nathan LaPierre, Fanny Chevalier, and Derek Reilly. 2013. Tweetris. *Proceedings of the 9th ACM Conference on Creativity & Cognition*

- *C&C '13*: 224.

13. Uta Hinrichs, Simon Butscher, Jens Müller, and Harald Reiterer. 2016. Diving in at the Deep End: The Value of Alternative In-Situ Approaches for Systematic Library Search. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*: 4634–4646.

14. Eva Hornecker and Emma Nicol. What Do Lab-based User Studies Tell Us About In-the-Wild Behavior ? Insights from a Study of Museum Interactives.

15. Rose Johnson, Yvonne Rogers, Janet van der Linden, and Nadia Bianchi-Berthouze. 2012. Being in the Thick of In-the-wild Studies: The Challenges and Insights of Researcher Participation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*: 1135–1144.

16. Jesper Kjeldskov and Mikael B Skov. 2014. Was it Worth the Hassle ? Ten Years of Mobile HCI Research Discussions on Lab and Field Evaluations. *Acm*: 43–52.

17. Jesper Kjeldskov, Mikael B Skov, Benedikte S Als, and Rune T Høegh. 2004. Is It Worth the Hassle ? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. *Mobile Human-Computer Interaction Proceedings of the 6th International Symposium*: 61–73.

18. Jean Lave, Lucy Suchman, and Ed Hutchins. 2013. Introduction to the Special Issue of " The Turn to The Wild ." 20, 3: 0–3.

19. P Marshall, Y Rogers, and N Pantidi. 2011. Using F-formations to analyse spatial patterns of interaction in physical environments. *Cscw 2011*, Cscw: 3033–3042.

20. Paul Marshall, Richard Morris, Yvonne Rogers, Stefan Kreitmayer, Matt Davies, and Milton Keynes. 2011. Rethinking " Multi-user ": An In-the-Wild Study of How Groups Approach a Walk-Up-and-Use Tabletop Interface. 3033–3042.

21. Paul Marshall, Richard Morris, Yvonne Rogers, Stefan Kreitmayer, and Matthew Davies. 2011. Rethinking "Multi-User" - An In-The-Wild Study of How Groups Approach a Walk-Up-and-Use Tabletop Interface. *Proceedings of the International Conference on Human Factors in Computing Systems (CHI'11)*: 3033–3042.

22. Daniel Michelis and Jörg Müller. 2011. The Audience Funnel: Observations of Gesture Based Interaction With Multiple Large Displays in a City Center. *International Journal of Human-Computer Interaction* 27, 6: 562–579.

23. Peter Peltonen, Esko Kurvinen, Antti Salovaara, Giulio Jacucci, Tommi Ilmonen, John Evans, Antti Oulasvirta, and Petri Saarikko. 2008. ``It's Mine, Don't Touch!'': Interactions at a Large Multi-Touch Display in a City Centre. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*: 1285.

24. Derek Reilly, Fanny Chevalier, and Dustin Freeman. 2014. Blending Arts Event and HCI Research. In *Interactive Experience in the Digital Age: Evaluating New Art Practice*. 22–23.

25. Julie Rico and Stephen A. Brewster. 2010. Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*: 887–896.

26. Y Rogers. 2011. Interaction design gone wild: striving for wild theory. *Interactions* 18, 4: 58–62.

27. Yvonne Rogers, Kay Connelly, Lenore Tedesco, William Hazlewood, Andrew Kurtz, Robert E. Hall, Josh Hursey, and Tammy Toscos. 2007. Why It's Worth the Hassle - The Value of In-Situ Studies When Designing Ubicomp. *UbiComp'07* 4717: 336–353.

28. Jiamin Shi and Florian Alt. 2016. The Anonymous Audience Analyzer – Visualizing Audience Behavior in Public Space. *CHI Extended Abstracts on Human Factors in Computing Systems*: 3766–3769.

29. Robyn Taylor, Guy Schofield, John Shearer, Jayne Wallace, Peter Wright, Pierre Boulanger, and Patrick Olivier. 2011. Designing from within: humanaquarium. *Proceedings of CHI 2011*: 1855–1864.

30. Robyn Taylor, Guy Schofield, John Shearer, Peter Wright, Pierre Boulanger, and Patrick Olivier. 2014. Nightingallery: theatrical framing and orchestration in participatory performance. *Personal and Ubiquitous Computing* 18, 7: 1583–1600.

31. Daniel Vogel and Ravin Balakrishnan. 2004. Interactive Public Ambient Displays: Transitioning from Implicit to Explicit, Public to Personal, Interaction with Multiple Users. *UIST '04: Proceedings of the 17th annual ACM symposium on User interface software and technology* 6, 2: 137–146.

32. Annika Waern. 2016. The Ethics of Unaware Participation in Public Interventions. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*: 803–814.

33. Robert Walter, Gilles Bailly, and Jörg Müller. 2013.

StrikeAPose: revealing mid-air gestures on public displays. In *SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*, 841–850.

34. Julie R. Williamson and Daniel Sundén. 2015. Deep Cover HCI. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '15*: 543–554.

35. Julie R. Williamson, Daniel Sundén, and Jay Bradley. 2015. GlobalFestival: Evaluating Real World Interaction on a Spherical Display. *Proceedings of the Joint International Conference on Pervasive and Ubiquitous Computing and the International Symposium on Wearable Computers (Ubicomp/ISWC'15)*: 1251–1261.

36. Julie R. Williamson and John Williamson. 2014. Analysing Pedestrian Traffic Around Public Displays. In *Proceedings of The International Symposium on Pervasive Displays - PerDis '14*, 13–18.