In-the-moment and beyond: Combining post-hoc and real-time data for the study of audience perception of electronic music performance

S. M. Astrid Bin, Fabio Morreale, Nick Bryan-Kinns, and Andrew P. McPherson

Centre for Digital Music, Queen Mary University of London, London, UK
a.bin, f.morreale, n.bryan-kinns, a.mcpherson @ qmul.ac.uk

Abstract. This paper presents a methodology for the study of audience perception of live performances, using a combined approach of post-hoc and real-time data. We conducted a study that queried audience enjoyment and their perception of error in digital musical instrument (DMI) performance. We collected quantitative and qualitative data from the participants via paper survey after each performance and at the end of the concert, and during the performances spectators were invited to indicate moments of enjoyment and incidences of error using a two-button mobile app interface. This produced 58 paired post-hoc and real-time data sets for analysis. We demonstrate that real-time indication of error does not translate to reported non-enjoyment and post-hoc and real-time data sets are not necessarily consistent for each participant. In conclusion we make the case for a combined approach to audience studies in live performance contexts.

Keywords: Audience studies; error; performance studies

1   Introduction

As computing has found its way into every facet of our lives, the experience of the user has become a central point of HCI study and discussion [38]. Indeed, as interfaces have become ubiquitous and their applications have expanded to include a broader and more personal range of interactions - far beyond task-based interactions in the workplace [27] - their use has become ever more ambiguous, and their purpose potentially open to many interpretations [32].

Experimental electronic music (EEM) performance using digital musical instruments (DMIs) shares many of these same features of ambiguity. EEM evolved out of the 20th century avant-garde and valued experimentation and improvisational exploration over musical vernacular. Because of this emphasis on improvisation, in EEM performance there is no specific task and no "right" or "wrong" interaction - a commonality it shares with the current HCI paradigm. (It should be stressed that in this sense we are concerned with DMIs as tools of improvisation, and not in the details of their usability.) With decades of development,

EEM performance has been exploring ambiguous interactions in an audience context for far longer than third-wave HCI, and we propose that it is a fertile ground for inquiry into spectator experience of ambiguous interactions.

In HCI audience studies, and certainly those in the New Interfaces for Musical Expression (NIME) research community, post-hoc methods are common means of data collection [3, 6, 8, 15, 21, 23]. Using surveys and/or interviews, investigators can quickly and inexpensively gather a wealth of quantitative and qualitative data based on audience opinions. However, Loftus and her collaborators demonstrate that human memory is notoriously unreliable [25], which raises questions about, if not the veracity of post-hoc data, then what additional conclusions real-time data may allow us to draw, and what other methodologies might be employed.

We were therefore motivated to investigate the role and content of both post-hoc and real-time data in an audience study, in order to develop a methodology that might make the best use of both. The questions we explore in this paper are as follows:

1. What kind of evaluation can be undertaken with post-hoc and real-time feedback, and how can these forms of data collection inform one another?
2. What are the features of an incidence of audience enjoyment? What are the features of an incidence of "error"?

This paper presents a study that examined the role of familiarity and musical style in the enjoyment of DMI performance using this combined methodology. The study context was an evening concert, where two performers played self-built musical instruments in both an experimental and a vernacular style, and data was collected from the audience As well as post-hoc survey data, real-time data was collected via a custom-built system called Metrix, a system for collecting real-time audience feedback that runs on mobile phones and records spectator indications of "enjoyment" and "error" via a two-button interface.

The implications of the post-hoc data are discussed in depth in [2]. In this paper, we shift our focus to examining the results of combining of real-time and post-hoc data, and using these results to examine the specific notion of "error" in performance. We discuss the kind of evaluation that is possible with this combined methodology, and compare and contrast the post-hoc and real-time data sets – considering the how they inform one another, as well as the advantages and drawbacks of each. We also examine the real-time data to gain insight into the perception of the perception of EEM's ambiguous interactions, and performance features that may indicate "enjoyment" and/or "error", and how this might inform our understanding of each.

2   Related Work

In this section we first trace the history of real-time audience data collection, and existing methods for HCI studies of spectator perception. We also contextualise DMI performance in relation to third-wave HCI, and specify why error is an issue of interest in both arenas.

2.1 Data collection in audience studies

Real-time audience response has been measured since 1930s where it was first used to gauge audience response to radio, film, and television [26]. These studies took place in a lab, where spectators indicated their reactions with buttons and knobs on hand-held devices. Since then, real-time data gathering techniques have become more sophisticated and integrated into the performance setting, now including physiological data (such as head tracking [29] and galvanic skin response [24]), verbal and non-verbal feedback [1, 11], as well as the measurement of crowd behaviours, such as applause [4, 9].

Stevens et al. [34] describe a study done with pARF, a system comprised of 20 hand-held (PDA) computers programmed to gather time-series "arousal" data. Participants indicated their emotional state with a stylus on a 200px x 200px grid on the device's screen, and their response was measured at a rate of 2Hz. The devices were distributed to 20 individuals in an audience of 200 for feedback during a dance performance.

Though rigorous analysis of the real-time data gathered with pARF was performed, there are drawbacks to this method. First, the pARF system supports up to 20 devices (only 18 were used for the study), meaning that a 10% subset of the audience used it, a small sample that is generalised to a much larger crowd. Secondly, no post-hoc data was collected alongside pARF (except for demographic details), which also leaves open the question of the difference in insights this method might have when compared to post-hoc data. Metrix, the system we designed and used for our case study, addresses these gaps.

2.2 DMI performance, HCI, and the role of context

Though a comprehensive history of DMIs and EEM performance is beyond the scope of this paper, it is helpful to trace their roots. This musical tradition developed over the last 100 years (starting well before the advent of digital technology), and is connected to the avant-garde that rose out of seismic shifts in culture taking place in the early 20th century [18]. Connected to Russolo's Futurism [31] and timbre-focused work of Varèse, it emerged at a time of radical experimentation that made liberal use of new technology, and soon exposed the limits of the usual tools. As Varèse remarked, "Our musical alphabet must be enriched. We also need new instruments very badly ... which can lend themselves to every expression of thought and can keep up with thought" (1916, quoted in [37]).

Along with a pursuit of new instruments, practitioners set aside musical vernacular (features such as melody, triadic harmony, rhythmic regularity) in favour of radical experimentation. The path of development can be seen running thorough Pierre Schaeffer's *musique concrète* and the work of John Cage and his experimentation with the musical score. He created scores more akin to recipes, descriptions of musical situations that had to be produced and completed by the performer (and sometimes the audience).

This lack of established artistic goals and the discarding of established musical frames of reference parallels features of 3rd wave HCI [38]. HCI was originally concerned with task completion in the workplace [27], but since that time interaction with computers features in virtually every facet of life, and computing now serves much more nuanced social, emotional and cultural purposes. As such, "emotions and experiences are keywords in the third wave." [7] In these interactions, the task may be set by the user, the task may only become apparent during the interaction, or there may be no task at all. Gaver et al. [17] propose that this ambiguity in interfaces is a "resource for design" that, instead of leading users through a task, instead provides a space of possibility for interpretation. Further, Sengers and Gaver [32] assert that HCI "can and should systematically recognize, design for, and evaluate with a more nuanced view of interpretation in which multiple, perhaps competing interpretations can co-exist."

In both the DMI and HCI contexts, understanding what is (or isn't) done with the interface is as crucial as the device itself. Reeves et al. [30], suggest that the experience of an HCI spectator can be described by how they see the interaction, coupled with the effect of that interaction - whether the interaction and the outcome was hidden, partially hidden, transformed, revealed or amplified. In this way, the performer action and outcome are tightly coupled; more importantly, the notion of "task" is removed from the discourse.

Spectator experience of DMI performance is an area of interest in NIME, where discussion has settled around the notion of *transparency*. Transparency is defined by Fels et al. [12, 30] as "the psychophysiological distance, in the minds of the player and the audience, between the input and output of a device mapping" [12]. Since DMIs do not have to conform to traditional modes of interaction [20], considerable effort has been made to expose the instrument's workings to audiences, through visualisations of computational processes [5, 28] and physical metaphors [10].

Curiously missing from this discussion is the influence of musical style on audience perception. Whether an input-output mapping is understandable to the spectator may depend at least in part on whether they are *witting* spectators [33]; that is, whether they understand the norms of the musical style in which the instrument is used. Just as many interactions between humans and computers cannot be removed from their cultural context, instrumental transparency may only be measurable in the context of a particular musical style. This was a primary motivator for our case study, which questioned the role of familiarity on audience response to EEM performance using DMIs.

2.3 Perception of error in DMI performance

Fyans et al. have made inquiries into the spectator experience, particularly where it relates to the notion of error [13–16]. In one such study [15] they observe that spectators are able to identify few errors with DMI performance, even raising the question of whether error is even possible. Gurevich [19] contributes a more flexible system of thinking about boundaries and straying outside them,

by suggesting that variation is the locus of *style*, which he defines as individual variations.

It is important here to consider this notion of error. From the Latin *errare*, meaning *to stray*, "error" suggests a stepping outside of accepted boundaries. Kruse-Weber and Parncutt [22] define error in a classical music context as "unintended result of an action", and classify intended actions as those specified in the score. However, experimental electronic music performance is highly improvisational and has no vernacular resembling a classical score. In this context, how can a performer stray out of bounds? Are errors even possible?

This study on familiarity's impact on audience enjoyment presented an intriguing opportunity to also examine the notion of "error" in this context. We wanted to gain insight into whether enjoyment and error are mutually exclusive, and determine if features of these two states could be extracted from real-time indications by the audience. Therefore, our real-time system had buttons for indicating two states, "enjoyment" and "error".

Of course, this single audience study can't answer these questions in a general sense. It does, however, provide some intriguing insights that suggest what the audience perceives as errors as the performance unfolds.

3 Real-time data: Metrix

3.1 Motivation and technical description

When considering which system to use to collect real-time feedback, we first looked to existing solutions [1]. However, we found these solutions to be inappropriate for one or more of the following reasons: Prohibitively expensive; overly complicated; lacking in features (or drastically over-featured); hard to customise; requiring significant participant training; or generally unfit for purpose in this context. Leveraging the availability of web technologies and the ubiquity of personal smart devices, we designed Metrix, an application that was streamlined, easy to use, fit for purpose and customisable.

Metrix is an open-source system for real-time data collection. It is a single-page web app, and is designed to be used on mobile phones. Metrix runs on a web server, and users connect to it via their phone's browser. When active, participants can tap the interface's buttons, and the system records each user's button taps (grouped by the button that produced them) in a database as time stamps associated with their username. The resulting timestamp data can then be distributed along a timeline. (Figure 1 describes the Metrix dataflow).

The data gathering interface consists of a screen split in half into two buttons (Figure 2). It is inactive until the start of the performance, when it is made active by an investigator via a remote control interface.

There are significant benefits to this web app approach. First, this app runs on a mobile phone and there is nothing to download, meaning that an entire audience can participate (in similar studies, devices were custom and limited

---

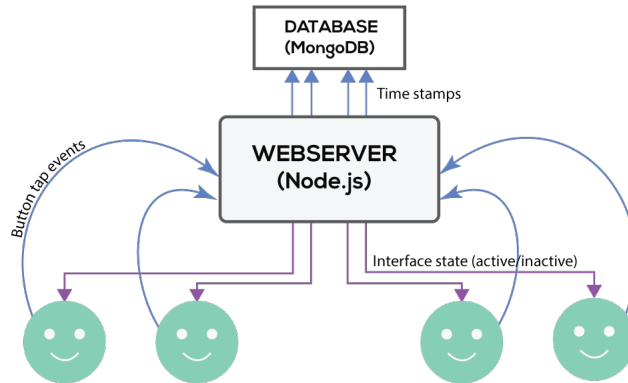[1] http://www.pcipro.com/, https://www.feedbackr.io/, http://www.clikapad.com/

Fig. 1. Diagram of data flow between participants and Metrix.

and only a small percentage of the audience could participate [35]). Second, the design of the interface is a web page, and is therefore easily customisable and can go through multiple design iterations. Third, this system leverages the ubiquity of mobile phone technology; audience members in many contexts can be assumed to have their own devices that they are already know how to use, so there is little on-boarding necessary.

Additional features such as username assignment also allows connections to be made in the datasets. When a participant accesses the Metrix interface, a user-name is automatically generated for them (an amalgamation of two randomly-chosen words) and displayed on their screen, and they are reminded of this username whenever the interface is inactive between performances.

There are, of course, contextual considerations when implementing Metrix as a research tool. Though we were leveraging the ubiquity of mobile phones in the context of our study, not all audience members everywhere will have a mobile phone. Additionally, a wifi connection that can support all users is needed, and web server load would be a consideration for very large audiences. Mobile phone batteries are also a factor, as audiences will probably not arrive with their phones fully charged, and Metrix requires that they be active for the entire performance.

Interface and interaction design We chose a two-button interface for Metrix. We were interested in the audience indicating two states: "I am enjoying this" and "There was an error" (we will hereafter refer to these buttons as "enjoyment" and "error").

Since this is an interface designed to be used during a performance, we wanted it to be as easy, intuitive, and unobtrusive as possible. A slider with a neutral position, for example, may require visual attention, so we opted for discrete buttons. The active interface is split in half, each half serving as a "button", so it was easy to tap each side without having to look at the device. We also chose not to use text to reduce cognitive load, and instead used symbols and colours on the buttons to indicate their function. On the left, "enjoyment" is green and

Fig. 2. Views of Metrix in use. Clockwise from top left: Screen to select group; presentation of username; view during data collection; post-performance pause and reminder of username.

is indicated by a :) symbol. On the right, "error" is red and is indicated by an X (see Figure 2). Each button provided some subtle feedback by darkening slightly when tapped.

4 Case Study

The context of this study was an evening concert, during which two musicians - each of whom plays a self-built DMI - gave two short performances of approximately five minutes each: one in a highly experimental style, and one in a conventional (vernacular) style. In this section, we will detail the study method.

4.1 Post-hoc data collection: Surveys

The audience for this study (N=64) was randomly distributed survey booklets on arrival. The book a participant received placed them in either Group 1 or Group 2. The survey booklets contained four short surveys to be completed after each performance (the *post-performance* surveys) and a longer survey to be completed at the end (the *post-concert survey*)[2]. The post-performance surveys asked three quantitative (rating scale) questions, and three qualitative (open-ended) questions. The post-performance survey asked more reflective questions and gathered demographic data. These survey answers were matched to the participants' real-time data through the Metrix username, which we asked our participants to write on the front of their survey books.

---

[2] Blank questionnaires available at: http://bit.ly/1QYBlIk
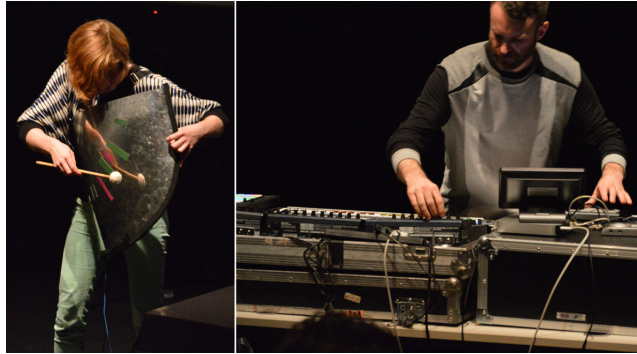
Fig. 3. The study performers, from left: Dianne Verdonk on La Diantenne; Tim Exile on the Flow Machine.

4.2 Study Design

The two musicians recruited for this study were Dianne Verdonk on the La Diantenne [36], and Tim Exile on the Flow Machine[3]. These musicians were chosen because they have achieved a level of virtuosity with their instruments, their instruments allow them to play in both an experimental and conventional musical styles, and the way their instruments work is not already familiar to an observer.

The audience was divided into two groups, according to which survey book they were handed upon entry. Group 1 received a ten-minute technical tutorial on Tim Exile's Flow Machine, and Group 2 received a ten-minute technical tutorial on Dianne Verdonk's La Diantenne.[4] This was to provide a difference of familiarity – for each performance, one group would be familiar with how the instrument worked, and the other would be unfamiliar.

While each group received their instrument tutorial, the other group received a short (10 minute) on-boarding session in another room that featured a two-minute video on how to use Metrix, and left time for questions. During this session we stressed that the states indicated by the buttons were not opposite - it was not "I heard an error" and "I didn't hear an error", or "I'm enjoying this" and "I'm not enjoying this". We also stressed that participants could tap the buttons as often - or as rarely - as they wished, and that the boundaries of "enjoyment" and "error" were entirely up to them.

The concert consisted of each performer playing two pieces, one experimental and one vernacular (the performers were asked to interpret this in the context of their individual performance practice). The order of performance was as follows:

---

[3] http://techcrunch.com/video/create-live-edm-with-tim-exiles-flow-machine/519373211/

[4] These tutorials were given by a member of the research group to avoid any bias toward seeing a particular performer before the show.

1. Dianne: Experimental
2. Tim: Experimental
3. Dianne: Conventional/Vernacular
4. Tim: Conventional/Vernacular

5 Processing of Real-Time Results

Prior to analysis, some data cleaning techniques were applied to the real-time data set. These included:

Truncating all tap events to the nearest second: The time stamps collected were in milliseconds, but that resolution of time proved too noisy. For that reason, all tap events were grouped by the second in which the tap took place.

Grouping taps by time interval: We grouped the "error" events by 1-second interval, and grouped the "enjoyment" events by 5-second interval (the reasoning for this is discussed in Section 5.2).

Filtering to remove multiple taps from intervals: A small number of participants appeared to be very enthusiastic with their button tapping, and tapped many times in a given time interval. To avoid one person's repeated tapping creating an artificial spike in tap events, we counted only one tap per participant in any given interval.

5.1 Data considerations

In a previous paper [2] we considered the post-hoc results and what they revealed about the effect of instrument familiarity and musical style on audience enjoyment of experimental DMI performance. For this analysis, we will instead focus on the real-time data in order to determine the features of enjoyment and error, and make reference to the post-hoc ratings of Enjoyment as we examine how these datasets complement one another.

It should be noted that participants were under no obligation to take part in both methods, and some only took part in one. We collected 64 surveys, but for this analysis have only included survey data that had a real-time data set from the same participant (58 participants in total; Group 1 n = 30, Group 2 n = 28).

It should be highlighted that since the notions of "enjoyment" and "error" were not considered to be complementary ideas, they were treated as different data sets with separate insights, and coded by the investigators entirely separately.

5.2 Process of Analysis

The first step was visualisation of the real-time data in histograms, using the 1 second bin width. For the histograms associated with the "error" button, the results were understandable at this time resolution. However, for the "enjoyment"

data, a 1s bin meant the data was still very noisy. The bin size was increased, and at 5s peaks became more prominent. An example of the distribution of "enjoyment" indications throughout a performance is illustrated in Figure 4.

5.3 Video coding

Two investigators independently analysed the histograms for "enjoyment" and "error" for each performance. The performances were recorded on video, with an audible click to mark the point where the Metrix interface was made active by the investigator. This made it possible to sync the video footage and the real-time data, which enabled us to analyse the performance and look for events in the performance around points of audience agreement about "enjoyment" and "error".

In coding the video, we defined what constituted an "event" for the video analysis as agreement among 3 or more people in one or two consecutive time intervals, preceded by two or more seconds of zero error indications. (See Figure 5 for an illustration.) The reason for two seconds of no indications was so we could be sure the previous event had ended.

6 Findings

After the video coding was completed by two investigators, a deductive thematic analysis was performed to extract themes, combining the results until saturation.

6.1 Features of "error" events

We found that audience-indicated "error" events were, for all performances, less common than "enjoyment" events. These "error" events tended to occur together across the audience, appearing as spikes in the histograms (see Figure 8, Note 2), whereas "enjoyment" events tended to occur far more often but with less agreement among the audience.
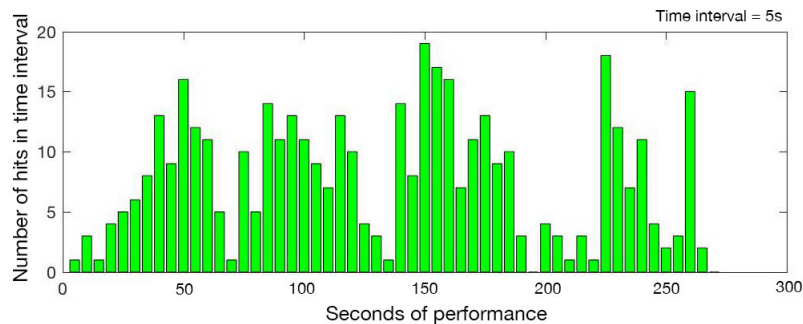
Fig. 4. Patterns of use for "enjoyment" button by 5s interval, Performance 1.

From our video coding and thematic analysis, we found that error events fell into the following categories:

1. Obvious and trivial performer error (Dianne at one point hit the mic stand, and Tim's rig shut off at the end of his final performance, which were widely indicated);
2. Sounds that were loud, or unexpected;
3. Facial expressions indicating a mistake;
4. Errors in musical content (for example, out of tune or against the expected rhythm).

In the experimental performances, the errors were primarily in categories 1, 2 and 3. In the conventional performances, errors in category 4 were also observed, suggesting that the audience had a deeper knowledge of the musical style in these cases.

6.2 Features of "enjoyment" events

"Enjoyment" events were not as straightforward as "error" events. Instead of appearing as spikes in the data, their appearance resembled a Gaussian distribution; taps increased to a peak over time, and then tapered off (see Figure 4 bottom for an example). "Enjoyment" appears to have a slower, more cumulative effect, contrasting the "error" events' sudden onset and sharp drop off.

Enjoyment events clustered around events with the following features:

1. Moments of novelty, such as the introduction of a new playing technique, timbre or texture;
2. Moments of high musical intensity, flow, or complex rhythmic patterns.

In the experimental performances, category 1 (novelty) was the driving factor in periods of enjoyment. In the conventional performances, both categories were observed, but category 2 (intensity) predominated. This again suggests an audience engagement with the underlying musical language.
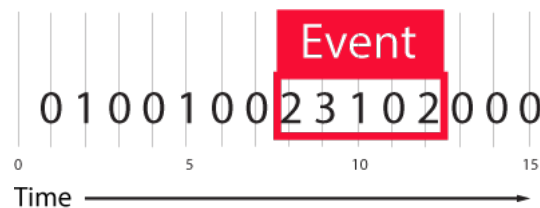
Fig. 5. An illustration of an "error" event, by examining the number of participant indications per second.

6.3 Real-time data compared to post-hoc findings

We compared the number of button taps (which we refer to as "indications") during the performances to see if an increased amount of "enjoyment" or "error" indications (Figure 9) bore any resemblance to the rank ordering of Enjoyment from the post-hoc data. This ranking - comparing those who were familiar with the instrument vs those who were unfamiliar with the instrument for each performance - is illustrated in Figure 6.

Considering the four performances overall, increased use of the real-time "enjoyment" button did not correlate with the audience's post-hoc rankings of enjoyment. Although the most "enjoyment" indications did occur during the highest-ranked performance (P4), that was the only similarity. There was no clear relationship between number of "error" indications and rankings of the performances: the lowest-ranked performance did have the most "error" indications, but this pattern did not hold for the other three performances.

6.4 Inconsistency between post-hoc and real-time reporting

In the post-performance surveys filled out immediately after each of the 4 performances, we asked participants to rate how enjoyable they found the performance they had just seen. We found inconsistency between their real-time reporting and their post-hoc reflections. We will consider two illustrative examples:

One respondent from Group 1 (who saw Tim's instrument before the concert) made 137 "enjoyment" indications during Performance 4 (Tim's conventional piece), more than twice as much as any other performance (this was after eliminating more than one indication per second). But, in the qualitative assessment of the performance they reported *It was a bit flat.*

A respondent from Group 2 (who saw Dianne's instrument before the concert) made 108 "enjoyment" indications during Performance 4 – also more than twice as much as any other performance. In the qualitative feedback they reported that *It seemed a bit disjointed.*
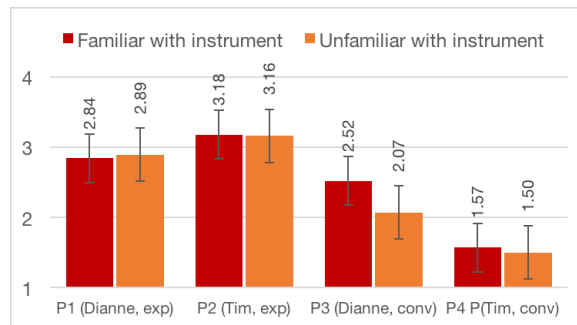
Fig. 6. Rank ordering of performances from post-concert survey, from favourite (1) to least favourite (4). Note that shorter bars indicate stronger preference.

		Enjoyment Indications		Error Indications	
		r value	p value	r value	p value
P1:	Familiar	0.37	0.05	-0.11	0.58
	Unfamiliar	0.18	0.34	0.36	0.36
P2:	Familiar	0.58	0.0007	0.19	0.32
	Unfamiliar	0.43	0.10	-0.32	0.10
P3:	Familiar	0.08	0.67	-0.16	0.42
	Unfamiliar	-0.11	0.57	-0.19	0.31
P4:	Familiar	0.27	0.15	0.09	0.63
	Unfamiliar	0.13	0.51	0.13	0.51

Fig. 7. A summary of the correlations of real-time "enjoyment" and "error" indications with the post-hoc rankings of Enjoyment of each performance.

These two examples suggest that raw numbers of real-time events are not always a good predictor of post-hoc reporting. More importantly, this inconsistency points to an intriguing area of study in audience perception: It suggests that what we think in the moment and what we think upon reflection may not be the same, and supports the need for examination of both post-hoc and real-time data as well as a way to better understand it.

### 6.5 Correlation of real-time data and post-hoc ratings

In our post-performance surveys, we asked respondents to rate the performance they had just seen according to the following questions:
1. How much did you enjoy the performance?
2. How interesting was the performance?
3. Did you understand how the instrument worked?

Since we are comparing real-time indications of Enjoyment, we will consider only the post-hoc Enjoyment rankings associated with Question 1 above.

For each performance and with each Familiar and Unfamiliar audience subgroup, we compared the rate of real-time "enjoyment" and "error" indications with the post-hoc quantitative ratings of Enjoyment, to see if there was any relationship (16 correlations in total, summarised in Figure 7). Across these, we found one correlation that was statistically significant. This correlation was between real-time "enjoyment" indications and post-hoc ratings of enjoyment (r=0.58 p=0.0007).

We found no positive or negative correlations between "error" indications and post-hoc rankings of enjoyment.
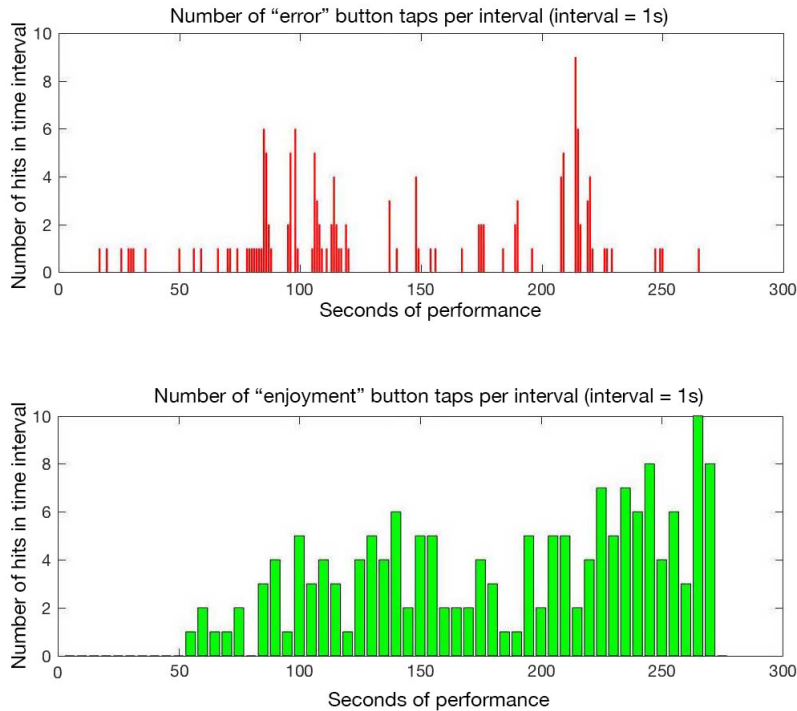
Fig. 8. Patterns of use of the "error" button and "enjoyment" button for Performance 3. NOTE 1: Error and good events often occur together, suggesting there is no binary relationship. NOTE 2: "Enjoyment" events are cumulative and rise to a peak, where "error" events have a sudden onset and sharp drop off.

7 Discussion

7.1 A consideration of the limitations of Metrix

The post-concert survey provided space for feedback on Metrix. Though feedback was overwhelmingly positive (with most comments referring to it as *easy* and *intuitive* and only one participant indicating that they found it *a little distracting*), we do acknowledge possible limitations of some features, and present data about why we believe these limitations do not impact the data we collected.

1. **Suggestion of binary states.** There is a risk that the two buttons, using complementary colours, suggest that the "enjoyment" and "error" states are opposite and mutually exclusive. Though we mentioned in the onboarding session that this is not the case, we acknowledge that simply telling an audience what they mean may not be enough.

However, our data showed that these buttons were not used in opposing ways, based on two observations. First, we found that there was no usage pattern that suggested that when the audience was not hitting the "enjoyment" button they

Avg per sec.	P1	P2	P3	P4
Enjoyment	1.567	1.227	0.600	1.893
Error	0.224	0.549	0.516	0.506

Fig. 9. Average number of "enjoyment" and "error" indications per second for each performance.

were hitting the "error" button, suggesting that they were not using them to indicate binary states. (See Figure 8, NOTE 1 for an illustration.)

Secondly, as we will describe in Section 7.2, the usage patterns of the "enjoyment" button and the "error" buttons was very different. "Enjoyment" events appeared to be cumulative events, gathering to a peak and tapering off again, whereas "error" events were sharp spikes with sudden drop off. This further suggests that the way the buttons were used was not related, and therefore our concern about a binary interpretation was unfounded.

2. **Use of symbols.** In order to reduce cognitive load we used symbols to indicate the two buttons instead of words: A smiley face indicated "enjoyment" and an X indicated "error". These were chosen because they were not similar symbols that could be confused, and did not suggest a binary relationship (as could be the case with a happy face and a sad face). However, we were concerned that these may not communicate clearly enough, and that respondents might still perceive a binary relationship.

Instead, we found, as described in point 1, that there didn't seem to be a binary relationship evident in the patterns of button use. There was also no feedback about the symbols being confusing. Further, a subsequent user workshop on this interface design with people who had never seen Metrix before indicated they found the symbols clear.

7.2 Observations on Usage Patterns

Over the course of the performance, there was no reduction in the use of the app; in fact, the final performance had the most indications on both buttons. We can thus conclude that Metrix kept participants engaged throughout. We cannot say specifically why this was, though simplicity or novelty may be factors.

We also found that participants vary widely in their willingness to record events, but that they appear relatively consistent in their level of use. We observed that those who tapped the button frequently for one performance tended to have similar levels of use for all four, and those who were sparing with their indications were similarly consistent. There will always be those who will be enthusiastic users and it is important to control for this in the analysis (for, example, by disregarding duplicate indications in a time interval).

7.3 Delayed reactions: Real time is still not immediate

Error and enjoyment events exhibited different temporal profiles. "Error" indications tended to happen in a narrower window, in clear response to a spe-

cific performer action or sound event. "Enjoyment" indications, however, appear spread out over wider time intervals. Clusters of "enjoyment" indications tended to build up over tens of seconds, reach a peak and taper off again, suggesting it is more of a persistent state than a discrete event.

This is not to suggest that "error" is a specific stimulus that gets a consistent and immediate reaction, and it is important not to view this as an action/reaction relationship. However, these differences do lend insight into the audience experience of both - that "error" is swiftly judged, whereas "enjoyment" tends to accumulate over time.

It is also challenging for participants to judge, and then register, a passing event that is usually part of the general continuum of a performance experience. As one participant noted, "*Enjoyable moments - as well as errors - pass so quickly. It took quite some mental time to decide and process, so sometimes I did not press at all because I felt the moment had already passed and it would not make sense at all to press anymore. The delay leads up to a sound, I would say, between the occurrence of an event and my button press.*"

7.4   A lot of errors doesn't mean that it's bad

The post-hoc rank ordering of the performances made a clear indication that Performance 4 was heavily preferred (see Figure 6). This, interestingly, was the performance that also got by far the most "error" indications in the real-time data.

This means that the presence of "error" indications did not suggest a performance that was not enjoyed. The suggestion here is that "error" perhaps does not deserve its negative connotations, and that there may be more to error than simply being something to avoid. It also suggests that events which audiences understand as "errors" are not necessarily bad; it is even possible that noticing subtle errors implies a certain level of audience engagement with a performance.

This finding is supported by the lack of negative correlation between "error" indications and ratings of Enjoyment both audience subgroups and all performances. If a tap on the "error" button suggested that the spectator considered the performance to be bad we would expect to find more widespread negative correlations between the number of these indications and the post-hoc ratings. This was not the case.

We found a positive correlation between "enjoyment" indications and ratings of Enjoyment for Performance 2, Tim's experimental performance, among both those familiar with the instrument. This is notable because this was the lowest-rated performance among all audiences. It may suggest that when there is no common vernacular on which to rely then instrument familiarity is not a significant factor in whether or not a performance is enjoyable (a finding supported by the post-hoc analysis of this study [2]). However, because these data points are isolated, more in-depth study is needed in order to formulate any specific conclusions.

7.5 What real-time and post-hoc data have to offer one another

Real-time data affords us second-by-second insight into audience experience, but it does not provide any contextual insight. Conversely, post-hoc data provides detailed contextual and descriptive feedback, but we have no way of tying these to any specific event. Each, therefore, has the potential to provide what the other cannot. This is potentially very powerful, and suggests that there are dimensions of audience data available through combining these two techniques that are inaccessible when only one is used.

But, each method can only ask its own questions, and we must be careful about which conclusions to draw. To make meaningful conclusions, the questions asked by both techniques have to be designed to inform one another, and how to best formulate questions to get results that can be meaningfully related requires further study.

Further, as we demonstrated in Section 6.4, we found inconsistencies between the post-hoc and real-time data for individual participants, which suggests that this potential multidimensionality of a combined methodology is not straightforward. It does suggest, however, an intriguing direction for future study, investigating why we think one thing in the moment and another afterwards, and what this means about the way we perceive time-based events.

8 Conclusion: Implications for HCI

This paper presented a combined methodology of real-time and post-hoc data collection for audience studies. We presented Metrix, a mobile phone-based real-time data collection system, that features an easy-to-understand UI and anonymised user ID generation that allows individual real-time responses to be linked to post-hoc surveys. We present this method in the context of digital musical instrument (DMI) performance, but it is applicable to any HCI context involving audiences or spectators.

We view this examination of EEM performance using DMIs through the lens of third-wave HCI, as in both the relationship between humans and technology is often subjective, ambiguous or culturally-dependent. We found that, in the real-time data, the rankings of performances in this study is not is not reflected in the frequency of "enjoyment" or "error" indications in real time, and that a lot of "error" events does not mean a performance was not enjoyed. Furthermore, we find that individual audience members are not always consistent between real-time and post-hoc responses, and suggest that more study is needed to examine why this is the case. We also find that neither set of results is predictive of the other. Finally, we reflect on real-time and post-hoc data providing different insights and how these may be related to provide more powerful insights into audience perception.

Because each data collection technique provides a dimension that the other cannot, future HCI studies within and beyond the arts domain may benefit greatly from a combined approach, but further study is needed to determine how

to most effectively understand and combine them to provide multi-dimensional insights that cannot be gained by the use of one technique alone.

9   Acknowledgments

References

1. Agarwal, A., Meyer, A.: Beyond usability: evaluating emotional response as an integral part of the user experience CHI Extended Abstracts on Human Factors in Computing Systems (2009)
2. Bin, S. A., Bryan-Kinns, N., McPherson, A. P.: Skip the pre-concert demo: How technical familiarity and musical style affect audience response. Proc. NIME (2016)
3. Barbosa, J., Calegario, F., Teichrieb, V., Ramalho, G., McGlynn, P.: Considering Audience's View Towards an Evaluation Methodology for Digital Musical Instruments. Proc. NIME (2012)
4. Barkhuus, L., Jørgensen, T.: Engaging the crowd: studies of audience-performer interaction. ACM (2008)
5. Berthaut, F., Marshall, M., Subramanian, S., Hachet, M.: Rouages: Revealing the mechanisms of digital musical instruments to the audience. Proc. NIME (2013)
6. Bilda, Z.: Evaluating audience experience. Creativity and cognition studios (2006)
7. Bødker, S.: When second wave HCI meets third wave challenges. Proceedings of the 4th Nordic conference on Human-computer interaction (2006)
8. Bongers, B., Mery, A.: Interactive kaleidoscope: audience participation study. Proceedings of the 23rd Australian Computer-Human Interaction Conference (2011)
9. Bull, P., Noordhuizen, M.: The mistiming of applause in political speeches. Journal of Language and Social Psychology, Vol. 19 No 3. (2000)
10. Dahl, L., Wang, G.: Sound Bounce: Physical Metaphors in Designing Mobile Music Performance. Proc. NIME (2010)
11. Desmet, P.: Measuring emotion: Development and application of an instrument to measure emotional responses to products. Funology, Springer (2005)
12. Fels, S., Gadd, A,. Mulder, A.: Mapping transparency through metaphor: towards more expressive musical instruments. Organised Sound, Vol. 7 No. 2 (2002)
13. Fyans, A. C., Gurevich, M., Stapleton, P.: Where did it all go wrong? A model of error from the spectator's perspective. Proc. NIME (2009)
14. Fyans, A. C., Gurevich, M., Stapleton, P.: Spectator understanding of error in performance. CHI'09 Extended Abstracts (2009)
15. Fyans, A. C., Gurevich, M., Stapleton, P.: Examining the spectator experience. Proc. NIME (2010)
16. Fyans, A. C., Gurevich, M.: Perceptions of Skill in Performances with Acoustic and Electronic Instruments. Proc. NIME (2011)
17. Gaver, W. W., Beaver, J., Benford, S.: Ambiguity as a resource for design. Proceedings of the SIGCHI conference on Human factors in computing systems (2003)

18. Gluck, M.: Toward a Historical Definition of Modernism: Georg Lukacs and the Avant-Garde. The Journal of Modern History Vol 58 No 4 845–882 (1986)
19. Gurevich, M., Stapleton, P., Bennett, P.: Designing for style in new musical interactions. Proc. NIME (2009)
20. Gurevich, M., Fyans, A. C.: Digital Musical Interactions: Performer-system relationships and their perception by spectators. Organised Sound Vol 16 No 2 166–175 (2011)
21. Höök, K., Sengers, P., Andersson, G.: Sense and sensibility: evaluation and interactive art Proceedings of the SIGCHI conference on Human factors. in computing systems (2003)
22. Kruse-Weber, S., Parncutt, R: Error management for musicians: an interdisciplinary conceptual framework. Frontiers in psychology (2014)
23. Lai, C., Boverman, T.: Audience Experience in Sound Performance. Proc. NIME (2013)
24. Latulipe, C., Carroll, E. A., Lottridge, D.: Love, hate, arousal and engagement: exploring audience responses to performing arts. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2011)
25. Loftus, E. F., Palmer, J. C.: Reconstruction of automobile destruction: An example of the interaction between language and memory. Journal of Verbal Learning and Verbal Behavior (1974)
26. Millard, W. J.: A history of handsets for direct measurement of audience response. International Journal of Public Opinion Research (1992)
27. Norman, D. A.: The Design of Everyday Things. MIT Press (2013)
28. Perrotin, O., d'Alessandro, C.: Visualizing Gestures in the Control of a Digital Musical Instrument. Proc. NIME (2014)
29. Poole, A., Ball, L. J.: Eye tracking in HCI and usability research. Encyclopedia of Human Computer Interaction, vol 1, 211–219 (2006)
30. Reeves, S., Benford, S., O'Malley, C., Fraser, M.: Designing the spectator experience. Proceedings of the SIGCHI conference on Human factors in computing systems (2005)
31. Russolo, L.: The Art of Noises: Futurist Manifesto Audio Culture: Readings in modern music. Continuum International New York (2010)
32. Sengers, P, Gaver, W. W.: Staying open to interpretation: engaging multiple meanings in design and evaluation. Proceedings of DIS (2006)
33. Sheridan, J. G., Bryan-Kinns, N.: Designing for performative tangible interaction. International Journal of Arts and Technology (2008)
34. Stevens, C., Glass, R., Schubert, E., Chen, J.: Methods for measuring audience reactions. Proceedings of the Inaugural International Conference on Music Communication Science (2007)
35. Stevens, C. J., Schubert, E., Morris, R. H., Frear, M.: Cognition and the temporal arts: Investigating audience response to dance using PDAs that record continuous data during live performance. International Journal of Human-Computer Studies (2009)
36. Verdonk, D.: Visible Excitation Methods: Energy and Expressiveness in Electronic Music Performance. Proc. NIME 2015
37. Wen-Chung, C.: Open rather than bounded. Perspectives of New Music Vol 5 No 1 (1966)
38. Wright, P., Blythe, M., McCarthy, J.: User experience and the idea of design in HCI. International Workshop on Design, Specification, and Verification of Interactive Systems (2005)