A User Experience Review of Music Interaction Evaluations

Dom Brown  
University of the West of England  
Bristol, UK  
dom.brown@uwe.ac.uk

Chris Nash  
University of the West of England  
Bristol, UK  
chris.nash@uwe.ac.uk

Tom Mitchell  
University of the West of England  
Bristol, UK  
tom.mitchell@uwe.ac.uk

ABSTRACT
The need for thorough evaluations is an emerging area of interest and importance in music interaction research. As a large degree of DMI evaluation is concerned with exploring the subjective experience: ergonomics, action-sound mappings and control intimacy; User Experience (UX) methods are increasingly being utilised to analyse an individual’s experience of new musical instruments, from which we can extract meaningful, robust findings and subsequently generalised and useful recommendations. However, many music interaction evaluations remain informal. In this paper, we provide a meta-review of 132 papers from the 2014 – 2016 proceedings of the NIME, SMC and ICMC conferences to collate the aspects of UX research that are already present in music interaction literature, and to highlight methods from UX’s widening field of research that have not yet been explored. Our findings show that usability and aesthetics are the primary focus of evaluations in music interaction research, and other important components of the user experience such as enchantment, motivation and frustration are frequently if not always overlooked. We argue that these factors are prime areas for future research in the field and their consideration in design and evaluation could lead to a better understanding of NIMEs and other computer music technology.

Author Keywords
Evaluation, Methods, Meta-Analysis, User Experience (UX), Human Computer Interaction (HCI), User Studies.

ACM Classification
A.1 [Introductory and Survey]; H.5.5 [Information Interfaces and Presentation] Sound and Music Computing—Methodologies and Techniques; H.5.2 [Information Interfaces and Presentation] User Interfaces—Evaluation/Methodology.

1. INTRODUCTION
Evaluation has become the subject of important discussion and consideration in the NIME and wider computer music community, and has been previously described as the “holy grail of NIME research” [20]. Perhaps due to NIME’s historic connection to SIGCHI, the field has often looked to the Human-Computer Interaction (HCI) community for inspiration in evaluation methods and frameworks [22, 36]. While early NIME evaluation methods focused on task-based usability [36], there has been a shift towards a more subjective and experiential focus [16, 22, 34], inspired by User Experience (UX), a movement within HCI that focuses on the user’s subjective experience of an interaction with technology [23, 28].

Despite this, many music interaction evaluations remain informal, and do not adhere to any particular method, while the qualities evaluated has become diverse (Figure 1) [1]. However, NIME evaluations often correlate with themes of UX, with researchers studying phenomena like engagement [35], emotion [11] and interest [3] in a computer music context, and with a desire for generalisable results, the NIME community would benefit from evaluations following established methods.

Barbosa et al. [1] make a valuable contribution in their analysis of NIME papers, and provide an overview that highlights the diverse nature of evaluations used in NIME research. In contrast, we will analyse recent interaction evaluations published at the NIME, SMC and ICMC conferences from the perspective of UX, using an adaptation of the QUOROM method used by Bargas-Avila and Hornbæk [2]. By doing so, we intend to provide a fresh perspective of music interaction evaluations through the lens of a separate, but very much related, discipline, examining recent trends and identifying areas for future consideration in the design and evaluation of NIMEs and other music interaction technology.

2. BACKGROUND
Various evaluation methods have been put forward to evaluate digital musical instruments and other computer music applications. The most notable of which is that of Wanderley and Orio [36], who borrow from HCI and suggest using
musical tasks to quantitatively evaluate the usability of musical controllers. While usability is an important aspect of HCI research and a useful metric in some musical domains, users of a musical technology are often more interested in its capacity for expression, or whether it is engaging, enjoyable, or rewarding over whether it is easy to use. As such, there has been a move within the NIME community away from usability as an evaluation target and towards more subjective, experiential-based methods [34, 16, 22] inspired by UX, with the need for suitable evaluation metrics becoming an important area of discussion [12].

3. OBJECTIVES
This paper seeks to provide an alternative perspective of the music interaction evaluations taking place in the computer music community by analysing recent literature from NIME, SMC and ICMC involving empirical user-focused evaluations. Our review will focus on:

1. The stakeholders considered in the evaluations.
2. The dimensions of UX that are evaluated.
3. What participant tasks are used.
4. How data is collected.

By studying these papers through a UX perspective, established criteria in the field such as virtuosity and transparency [14] are not explicitly considered, as they do not fit within the definitions of the UX dimensions.

4. METHODOLOGY
Our method is drawn from an adaptation of the QUOROM method [29] used by Bargas-Avila and Hornbæk [2]. We have filtered our corpus as follows:

1. Identify sources.
   Source Selection: Conference proceedings of New Interfaces for Musical Expression (NIME), Sound and Music Computing (SMC), and the International Computer Music Conference (ICMC) for 2014 – 2016 (N = 862).

2. Find appropriate publications.
   Screening criteria: Papers that mention an empirical user study in the title or abstract, using the keywords Evaluating, Evaluation, User, Study. (N = 147).

   Screening Criteria: Papers of which the evaluation focuses on the user’s experience (For example, papers that use musical Turing tests are omitted) (N = 132).

4. Final Corpus.
   The final corpus for our meta-analysis consisted of 132 papers.

For the UX dimensions, we chose to use similar dimensions to those found by Bargas-Avila and Hornbæk [2] in their meta-analysis to be prominent aspects evaluated in the UX field. We also decided to note when papers focus on usability, allowing us to compare its use against the dimensions of UX. We have provided the definitions used in categorising the corpus.

1. Usability Evaluations cover concepts such as ease of use, effectiveness and ergonomics, efficiency and learnability [30].
2. Generic UX Evaluations take a holistic approach and seek to explore the participants’ experiences as a whole, without focusing on any specific dimensions.
3. Aesthetics Evaluations focus on the aesthetic, artistic properties of the experience [24], such as appeal, taste, style, and expression [9].
4. Emotion Evaluations measure the emotional response and feelings of participants.
5. Enchantment Evaluations focus on the affective attachment of people to technology [25].
6. Engagement Evaluations study flow [8], intrinsic interest and curiosity [7].
7. Enjoyment Evaluations focus on the hedonic qualities of interaction [6].
8. Motivation Evaluations focus on the participant’s decisions and behaviour [13].
9. Frustration Evaluations focus on the participant’s dislikes and hindrances during an interaction [27].

In a similar method to Barbosa et al. [1], we have identified the stakeholders in each evaluation, using the following categories:

1. Performers Participants with agency, actively affecting their experience of real-time auditory interaction.
2. Audiences Participants without agency in the evaluation, passively involved in the experience.
3. Designers Participants with agency in evaluations that involve creating or designing hardware or software.
4. Composers Participants with agency in evaluations that involve composing or creating artistic material, but not performing.

We feel that it is important to give these definitions as some of the evaluations do not follow a traditional performance framework, for example in [17], where each participant is asked to play an auditory game. As the participant is actively engaging in a task, they have been categorised as a performer.

5. RESULTS
Due to our analysis taking place before the ICMC 2016 proceedings were available, the small number of relevant ICMC 2015 papers (N = 6) and the joint ICMC/SMC conference of 2014, we decided to group the ICMC and SMC papers together in our analysis. The breakdown of papers used in the analysis is as follows:

- NIME 2014: 28
- NIME 2015: 19
- NIME 2016: 15
- ICMC/SMC 2014: 32
- ICMC 2015: 6

Figure 2: The QUOROM procedure for this study.
Our analysis was non-exclusive, with some evaluations covering more than one UX dimension, data collection method, stakeholder or participant task. If more than one evaluations were included in a publication they were recorded as separate results. After our analysis, we identified the following categories for the participant tasks:

1. **Specific Task** Participants are asked to perform a pre-determined exercise, such as listen to auditory stimulus, or perform certain tasks with an instrument.
2. **Open Exploration** Participants are free to do as they please during an interaction.
3. **Guided Exploration** Participants have some freedom, but are guided by certain constraints.
4. **Watch Performance** Participants watch a performance given by a musician, in either a concert or laboratory setting (e.g. watching a video).
5. **Prepare and/or Give Performance** Participants are asked to prepare a piece and give a performance as part of the evaluation.
6. **Workshop** Participants’ interactions take place in a workshop setting.
7. **In The World Use** Participants use the technology in their own personal environments.
8. **Other** Any other task that does not fit in the above categories.

As well as data collection methods:

1. **Questionnaires** Specific questions used to gather responses.
2. **Likert Scales** Questionnaires use the Likert format.
3. **Comparisons** Participants are asked to compare stimulus, and give ratings; perform pair-wise comparisons and the like.
4. **Interviews** Either structured or unstructured.
5. **Field Notes** Observations are taken by researchers during the evaluation.
6. **Audio/Video Recording** Recordings of experiment are used in the analysis.
7. **Interaction Log** The user’s interaction with an interface is logged.
8. **Open/Informal Comments** Unstructured feedback is provided.
9. **Created Materials** Things made by participants during the evaluation are analysed, e.g. [26].
10. **Physiological Measurements** Methods such as EEG, ECG and the like are used to record a participant’s body.
11. **Other** Any other method that does not fit in the above categories.
12. **NS** The data collection method is not specified.

### 5.1 Stakeholders

The most popular stakeholder used in evaluations was the performer (50.7%), followed by the audience (39.3%). Designers (3.3%) and composers (6.7%) perspectives were rarely evaluated. While it has been suggested that performers are the most important stakeholders in digital music [4], our results suggest that the perspectives of designers and composers could be better represented during evaluations, as these perspectives may reveal aspects of musical interactions that have previously been overlooked.

Our stakeholders results are quite different to those of [1], whose stakeholders results were: Performers: 52, Designers: 28 and Audience: 20. We believe this is because of their inclusion of technical evaluations as evaluations from the designer’s perspective. Since our focus is on evaluations with participants our designers result is low, as a designer’s subjective experience is not usually solicited during technical evaluations.

### 5.2 UX Dimensions

Our results indicate that although UX concepts are being applied in computer music research, usability remains a popular metric in NIME papers (21.7%), while within ICMC and SMC, the largest proportion were not applicable to dimensions of UX, for example [15], in which an audience’s perception of vibro-tactile feedback is measured. A high amount of not applicable papers is to be expected, and is most likely due to the fact that empirical evaluations in computer music research do not always share the same targets as UX research, and so a large number of papers will not fit within our scope.

Of the dimensions of UX, aesthetics is the most commonly used (19.4%), followed by generic UX (13.7%) and engagement (10.9%). This reflects the literature of the field, which highlights the importance of expression [10], style [18, 19] and engagement [37] in computer music research. Generic UX papers often included evaluations with less formal structures, such as [21], in which a group of children are used to evaluate a museum experience through open exploration and group interview, and reflect the ideas of Stowell et al. [34] in their proposed qualitative method.

Emotion and enjoyment were evaluated in relatively equal measure (9.1%), but emotion evaluations in ICMC/SMC occurred only from the audience’s perspective.

Interestingly, three dimensions: motivation, enchantment and frustration; were evaluated for either rarely or not at all. This suggests that these are areas of UX that are currently overlooked in music interaction, and represent an opportunity for new directions in research. For example, studying how musicians become affectionately attached to an instrument may help us understand long term uptake of NIMEs, while studying motivation may allow us to explore their appeal over traditional instruments.

Although frustration is often linked to measurements of user error used in usability studies, in UX, frustration represents a qualitative exploration of negative aspects of a user’s experience, for example in [5], and its study could help the computer music community identify areas for improvement in the design of NIMEs and interaction technology.

#### 5.2.1 Performers

From the performer’s perspective, usability was found to be the most prominent dimension (29.4%), followed by generic UX (18.8%) while aesthetics, engagement and enjoyment share a similar proportion (10.5%). While NIME and
ICMC/SMC have different quantities of performer evaluations, they have a similar spread of evaluation dimensions, with usability being the most popular.

Usability remains prominent most probably because of its close relation to ideas of learnability and playability, which are important ideas in NIME and computer music research.

5.2.2 Audience

Aesthetics was the most prominent dimension from the audience’s perspective, in both ICMC/SMC and NIME. Interestingly, emotion was commonly studied within SMC and ICMC, while it was rare within NIME evaluations. Conversely, NIME often focused on engagement and enjoyment while ICMC/SMC evaluations rarely did so.

5.3 Participant Tasks

Overwhelmingly, the most popular participant tasks were specific tasks (53.1%), which make up the majority of ICMC/SMC evaluations. Meanwhile, NIME evaluations use specific tasks and open exploration in equal measure. The other tasks were used much less frequently.

Interestingly, NIME evaluations include watching performances much more than ICMC/SMC. This could be due to NIME’s focus on instruments, which suit audience evaluation through performance.

When filtered by UX dimension, it is interesting to observe that while questionnaires are the most popular technique for most dimensions, open exploration is the most popular for generic UX. This reflects the dimension’s less focused approach, in that via open exploration, any aspect of the interaction may be explored by participants. Similarly, “in the world” use is used mostly in generic UX, as this technique also encourages an open response from participants.

Meanwhile, emotion is studied nearly exclusively using specific tasks, with evaluations often asking audience participants to report on their emotions after listening to musical stimuli. The dimensions of aesthetics, engagement and enjoyment are each studied using a wide range of tasks, but most prominently specific tasks, open exploration and watching performances.

5.4 Data Collection

The most popular method of collecting data was by questionnaire (24.6%), and our results reflect those of [1]. Due to their prominence, questionnaires formatted as Likert scales were included in their own category (12.0%). Questionnaires most likely remain a popular technique as they give evaluations an ability to focus on specific aspects, and quantitatively analyse otherwise qualitative elements of an interaction.

Interaction logs are used mainly to measure usability. This reflects the evaluation technique of Wanderley and Orio [36], as well as Kiefer et al. [22], which use interaction logs to provide quantitative data for usability measurements.

Interviews and field notes were mostly used to measure aesthetic qualities.

While we found that emotional responses are elicited using specific tasks, they are collected using a wide variety of methods, including specific emotion measurement tools like the Self-Assessment Manikin (SAM).
This more qualitative perspective is also accounts and highlights the need for “rich personal experiences” of their previous use in HCI literature use descriptive methods, which may help us to design in ways that encourage new players; and en- courage us to design in ways that encourage new players. For example, looking at enchantment and the way in which musicians become emotionally attached to DMIs may help us to understand how short-term experimenters become long-term practitioners; understanding what motivates and influences musicians to choose DMIs could enable us to design in ways that encourage new players; and studying frustration in DMIs could help us to design more enjoyable and engaging music interaction experiences.

While these dimensions are inherently very different from each other, they share a very qualitative nature. Examples of their previous use in HCI literature use descriptive case studies and highlight the need for “rich personal accounts” [33]. This more qualitative perspective is also shared with much research in the NIME community, and highlights the growing trend in both UX and music interaction towards deeper explorations of a user’s subjective experience, as well as the potential ease with which these dimensions could be adopted into music interaction research.

As well as our UX dimension findings, we have found that specific tasks are the most popular participant task used in evaluations, and data is most commonly collected through questionnaires. While these are tried and tested methods, it indicates that there is room within computer music evaluations for the use of alternative methods, which may help us to evaluate our technologies more thoroughly.

For example, studying how musicians use instruments in their own personal environments (“in the wild”) allows us to better examine their creative process, as it is difficult to capture this in laboratory environments [16].

Similarly, the tasks of watching and preparing for a performance reflect real world use cases for musical technology, and we can learn much from studying the dynamics behind these processes. As every evaluation needs to be tailored to the specific goals and needs of the research in question [31], a full discussion of how our findings should affect future evaluations is beyond the scope of this paper, and is an area for future exploration.

Our analysis may have benefited from delineating between individual and group stakeholders, which would have provided a deeper insight into the user experience of multi-user interactions, such as collaborative installations. Also, breaking specific tasks into subcategories (for example into listening exercises and performance tasks) would have allowed for more detailed analysis of participant tasks.

By reviewing which areas of UX are commonly evaluated in music interaction research and which are overlooked, alongside the participant tasks and data collection methods used, we have provided a new perspective on the interaction evaluations taking place, and revealed alternative qualities to be considered in future NIME research.

6. DISCUSSION
Our results indicate that there is a strong correlation between UX and the evaluation criteria used in computer music research. However, usability remains the most prominent idea from HCI used in the field, despite efforts to move the field towards UX theories and principles.

We have found three common dimensions in UX research: motivation, enchantment and frustration; that are evaluated rarely or not at all in computer music interactions. These areas could help to address key questions regarding digital musical instruments, and help us to better understand the nature of the instruments and technologies we create. For example, looking at enchantment and the way in which musicians become emotionally attached to DMIs may help us to understand how short-term experimenters become long-term practitioners; understanding what motivates and influences musicians to choose DMIs could enable us to design in ways that encourage new players; and studying frustration in DMIs could help us to design more enjoyable and engaging music interaction experiences.

7. CONCLUSION
In this paper we have found that usability and aesthetics are commonly used in evaluations of interaction in the computer music field, while three areas of UX: motivation, enchantment and frustration; are often overlooked in current interaction evaluations, and represent potential avenues for future research. As well as this, we have found that questionnaires are the most popular method of data collection, and specific tasks are the most common participant tasks.

Future work will include the analysis of earlier years of NIME, SMC and ICMC to reveal how evaluations have evolved over time, as well as exploring how these findings may be applied to future evaluation methods.

8. REFERENCES


