

A Guide to Evaluating the Experience of Media and Arts Technology

Nick Bryan-Kinns and Courtney N. Reed

Abstract Evaluation is essential to understanding the value that digital creativity brings to people's experience, for example in terms of their enjoyment, creativity, and engagement. There is a substantial body of research on how to design and evaluate interactive arts and digital creativity applications Candy and Ferguson (2014). There is also extensive Human-Computer Interaction (HCI) literature on how to evaluate user interfaces and user experiences Blythe et al. (2004). However, it can be difficult for artists, practitioners, and researchers to navigate such a broad and disparate collection of materials when considering how to evaluate technology they create that is at the intersection of art and interaction.

This chapter provides a guide to designing robust user studies of creative applications at the intersection of art, technology and interaction, which we refer to as *Media and Arts Technology* (MAT). We break MAT studies down into two main kinds: *proof-of-concept* and *comparative studies*. As MAT studies are exploratory in nature, their evaluation requires the collection and analysis of both qualitative data such as free text questionnaire responses, interviews, and observations, and also quantitative data such as questionnaires, number of interactions, and length of time spent interacting. This chapter draws on over 15 years of experience of designing and evaluating novel interactive systems to provide a concrete template on how to structure a study to evaluate MATs that is both rigorous and repeatable, and how to report study results that are publishable and accessible to a wide readership in art and science communities alike.

Preprint. Chapter to appear in *Creating Digitally. Shifting Boundaries: Arts and Technologies — Contemporary Applications and Concepts*, Anthony L. Brooks (Editor), Springer.
<https://link.springer.com/book/9783031313592>

Nick Bryan-Kinns
Creative Computing Institute, University of the Arts London, London, UK, and
Queen Mary University of London, Mile End, London, UK e-mail: n.bryankinns@arts.ac.uk

Courtney N. Reed
Queen Mary University of London, Mile End, London, UK, and
Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany e-mail: creed@mpi-inf.mpg.de

1 Introduction

Media and Arts Technology (MAT) research exists in an interdisciplinary space at the intersection of artistic and creative practice, technology innovation, and research on human cognition and interaction. With MATs, the design of the creative application itself and the study of its use through user studies are intertwined and may take the form of a scientific intervention or an examination. This poses a challenge to you, as a researcher: on one hand, it offers opportunities for novel and engaging exploratory research and yet, at the same time, needs to be undertaken and framed in a way that works within current academic research discourse and vernacular. There is a natural tension between the focus of science and the arts when examined independently: science tends to focus on building and testing generalisable models and adding knowledge to our understanding of the world (England, 2016; Candy, 2011). Artistic practice approaches the understanding of the world slightly differently, often focusing toward creativity and the individuality and subjectivity of the human condition, whether or not that produces any particularly novel understanding of the world (Candy, 2011; Candy and Edmonds, 2018). However, these fields are intertwined and inseparable, having encouraged each other's advancement since the earliest ventures in philosophy and understanding of the world (Duarte et al., 2019; Benford et al., 2013). In order to conduct quality research in MAT, we must acknowledge both components: you as the researcher are responsible for honoring the inventiveness, innovation, and adaptability of the arts in a way which adheres to the structure and procedure of scientific research.

1.1 Principles of Quality Research

In order to make your work understandable to, and valued across research communities, it should meet two critical requirements: it must be both **rigorous** and **repeatable**:

Rigorous means that your research is conducted using tried and tested scientific methods and practices to collect and analyse data. MAT research is interdisciplinary, exploring new forms of digital Media and Arts, and yet is rooted in the science and engineering of the Technology. With these fields being extremely broad in their own right, MAT research must be diligent in conducting research that is appropriate for, and consistent with, existing bodies of knowledge. It is therefore vital to collect and analyse data in a structured way using established methodologies from scientific fields, to ensure that data collected and results analysed are considered to be valid and reliable by other researchers.

Repeatable means that your study plan and methodology are written in enough detail that someone else could run the study without you being present to explain it. This is important to maintain consistency of your study (each time you run the study it is done the same way), and to allow others to be able to reproduce your results if they want to.

Keeping these two core requirements in mind, you can design and evaluate compelling research which use MATs to contribute to the different fields in interdisciplinary research. The goal of this chapter is to outline practices for conducting and presenting research in a rigorous and repeatable way. In addition to providing guidelines for structuring your research, we illustrate these through examples of sound research practices used in existing MAT research.

This chapter starts by introducing the two types of MAT study — proof-of-concept and comparative — and provides example studies of both varieties, ranging from interactive music technology to playful tangible interaction. Referencing these existing projects, we introduce a series of guidelines for designing MATs as

part of an user study and explain how to define relevant research questions for your research. Then, we outline methods to determine appropriate data to collect from quantitative and qualitative sources and how these are combined as ‘mixed methods’. A guide to questionnaire and interview design is provided, which focuses on how to elicit further insights on experience from participants. Tools and techniques for qualitative and quantitative data analysis are introduced, including thematic analysis (Braun and Clarke, 2012) for interview data and statistical analysis for questionnaire and interaction data. The chapter concludes with guidance on how to report the results of the user studies so that others may be able to refer to your work and findings for their own design and technology.

In implementing the robust study design practices discussed in this paper, we hope that MAT researchers will find more common ground and understanding of each other’s multidisciplinary work. As such, this chapter serves as a guide for researchers from many disciplines studying MATs, leading to richer collaboration and dissemination of knowledge across research communities.

1.2 HCI, UX, and Interactive Arts

Before moving on, we would like to briefly discuss the historical link between the more computer science fields of HCI and user experience (UX) and digital arts, media, and creativity. The development of these fields has always been tightly intertwined (England, 2016; Nam and Nitsche, 2013), with the arts providing a source of inspiration for technology and creativity in its evaluation (Duarte et al., 2019). The earliest uses of computers to create art were in the 1950s, for example, in 1951 one of the earliest example of computer arts was the use of the Ferranti Mark 1 computer at the University of Manchester (there would only have been a handful of computers in the UK at the time) to play simple tunes. At a similar time Human Factors and Ergonomics, which are the origins of contemporary HCI and UX, emerged in the 1940s and 1950s to help designers design machines which were easier to use, for example, to improve the design of airplane controls to make flying safer and less error prone. The focus for Human Factors was really on the functionality of the machine. A lot of it was to do with how to layout, or configure, the controls in order to reduce chance of human error. This relationship is reciprocal and has worked in a collaborative way over the decades — typically, technology from other fields is adopted into arts, where it is used creatively. From these applications, new practices are developed and learned, stimulating the study and further expanding the technology itself. Interactive Art tends to place importance on what the user or listener cares about (Duarte et al., 2019; Jeon et al., 2019), as well as challenging the status quo to provide room for novel ideas (Benford et al., 2013). This provides a space for HCI to move beyond more traditional HCI topics such as task completion and efficiency to understanding and communication between humans and computational agents (Jeon et al., 2019).

2 Types of MAT Study

There are two broad types of MAT study which we will discuss in this chapter:

Proof-of-concept studies examine people’s responses to a single MAT. This kind of study asks “*What if...*” questions such as “What if I make this MAT, how do people respond to it, and what is their experience of it?” In this way, these studies often involve creating a MAT which is a digital intervention and then evaluating people’s response to it. These studies allow early-stage investigation of how people respond to a new form of

interaction, for instance when there is little existing research in the area and it is therefore unclear how people might respond to the interaction. Results of such a study can inform further studies by identifying broad kinds of response to the MAT and identifying possible interesting and challenging avenues for interaction design.

Comparative studies do as their name indicates – they compare two or more variations of a particular MAT. These studies compare the effects of specific design features and ask “*What effect do particular design feature(s) have on people’s experience?*” Rather than trying to determine whether one design is better than another, these studies are interested in *how* the experience is different between MATs. In this way, comparative studies involve in-depth examination of MATs and differences are typically restricted to one design feature. Results inform the development of theories about how design features might contribute to enjoyment, creativity, and engagement, and generate interaction design questions for future studies.

Often, proof-of-concept studies are used to research a general interaction question, which is then further refined and explored through specific questions in a comparative study. However, this is not necessarily the case and it should be made clear that one kind of study is not “better” or more rigorous or repeatable than the other – they simply ask different kinds of questions. As Section 3 will elaborate, it is important to keep your research goals in mind in order to decide and justify the methods you use.

In this chapter, papers published about existing MAT research projects are used to illustrate and bring to life these different study types, how they can be conducted, and how they are presented. These projects are briefly introduced in the following subsections to give a flavour for different kinds of MATs and user study approaches.

2.1 Example Proof-of-Concept Studies

The proof-of-concept studies we refer to throughout this chapter are very different from one another but share the same type of research focus and question: a MAT is designed in order to conduct an open-ended exploration of users’ behaviour and interaction. As mentioned above, although the research questions are very specific, they ask more of “*What if?*” or “*How do...?*” questions and are aimed at exploring the general effects of a MAT’s use as illustrated below.

2.1.1 Mazi

Nonnis and Bryan-Kinns (2019b) designed and studied Mazi, a Tangible User Interface (TUI) which uses haptic and auditory feedback to encourage spontaneous and collaborative play between children with high support needs. Mazi was developed through an iterative prototyping process and used in a proof-of-concept style study to explore how principles of TUI design along with theories of social interaction could be used to encourage social play (Nonnis and Bryan-Kinns, 2019a). Mazi’s final design features a dome-like shape to facilitate the circular configurations found naturally in communicative behavior and uses soft yet durable materials to allow the children to play in their own way with sensors embedded in Mazi which generate music. The core proof-of-concept research question of Mazi was:

“What if I make this MAT a large, soft circular shape and make it create music when multiple children play with it at the same time, how do children with autism respond to this, and what playful and social interaction does it prompt?”.

Because the authors wanted to determine how the tangible and auditory feedback of the design influenced the communication between the children, it was most suitable to conduct a proof-of-concept study; indeed, the wide variety of abilities and interests of children involved makes it unsuitable for comparative study. The proof-of-concept study took place over five weeks with five children aged between 6 and 9 years old at a Special Education Needs (SEN) in London.

The study process needed to be flexible to the needs of the children and teachers. In the proof-of-concept study, the researchers collected observational data about how the children interacted with each other and with Mazi, and analysed these using existing behavioural science models. The results of the study demonstrated that working with Mazi helped the children to master basic social skills and engage with different sensory interactions (Nonnis and Bryan-Kinns, 2019a). This open-ended strategy allowed the researchers to focus on the most salient elements of the interaction in-context, and yet at the same time produced a study method and results which could be replicated by other researchers.



(a) Mazi, made of wool and featuring inflatable bubbles for triggering sounds.



(b) Children playing together with Mazi.

Fig. 1: Mazi, a tangible user interface for stimulating interaction and participation for autistic children. Images used with permission from the Authors: <http://isam.eecs.qmul.ac.uk/projects/Mazi/mazi.html>

2.1.2 VoxEMG

Reed and McPherson (2020) design and explore the use of novel vocal interaction through surface electromyography (EMG) using an autobiographical approach (Neustaeder and Sengers, 2006). The authors developed a system, the VoxEMG, for gathering the electrical neural impulses which cause muscular contractions (Figure 2). These EMG signals are used in real-time sound design to allow a singer to interact with very low-level movements in their practice through auditory feedback. The auditory feedback allows the vocalists to interact with their existing, embodied understanding of their action and “hear” movements which would not normally

produce definable sound. The work was formed through an autobiographical approach in Reed's extended experience with the setup lasting over a year. The fundamental research question was:

"What if I sonify movements and actions which singers are not normally consciously aware of, and how will they react, change, and/or perceive their movements when they receive this new feedback?"

Autobiographical methods and first-person accounts are extremely useful as they demonstrate how lived-experience and understanding of a system's use can improve and inform its design (Höök et al., 2015). In this case, Reed applied her experience working as a semi-professional vocalist to the long-term interaction with the EMG system. Through detailed interaction notes, journaling, debugging, and an iterative design and testing process, Reed was able to uncover subtle understanding of the interactions with it (Neustaedter and Sengers, 2006; Reed and McPherson, 2021). This study again is more suitable as a proof-of-concept because it requires an exploratory approach; although there are some hypotheses as to how the musicians would respond, the researchers wanted to keep the interaction open-ended and see how the singer's behaviour changed while using this MAT.



(a) The VoxEMG board developed for sensing activation of the laryngeal muscles.
 (b) Reed wearing the Singing Knit wearable collar for vocal EMG interaction.

Fig. 2: Sensing and interacting with laryngeal muscular activations through VoxEMG integrated into wearable designs (reproduced from Reed et al. (2022)).

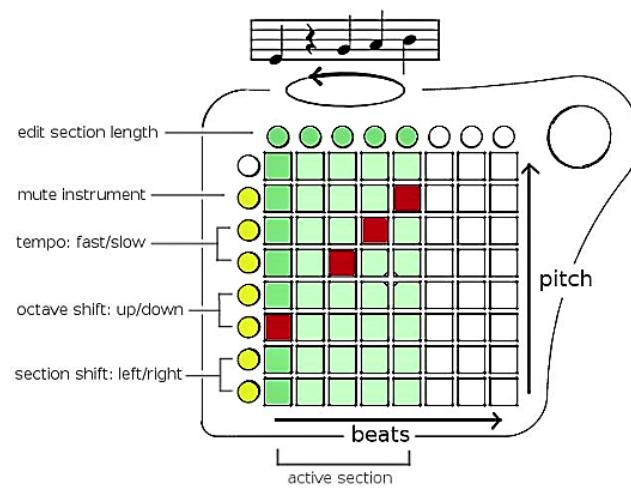
2.1.3 Polymetros

Polymetros is a collaborative music system designed as an in-person audience experience for multiple participants (Figure 3) (Bengler and Bryan-Kinns, 2013). The design is purposefully simple, using minimal music, and allowing participants to create short loops (8 notes long with 8 possible pitches and only one instrument sound). There are 3 physical instruments in Polymetros (Figure 3a) which are synchronised together so that the loops of each instrument are synchronised with each other. One person can play Polymetros on their own, but the sound becomes richer and more interesting as 2 or 3 instruments are played at the same time (Figure 3b). Essentially the research question of this MAT was:

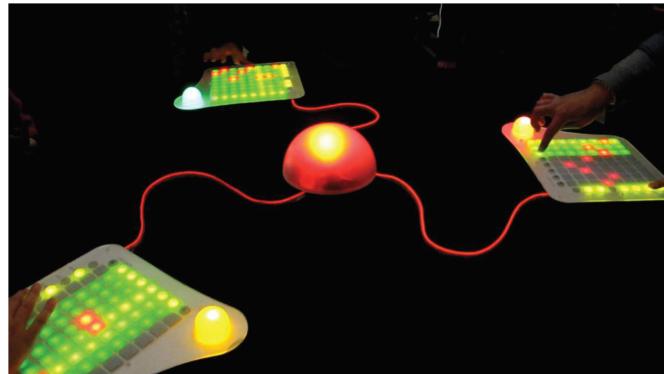
"What if I make a MAT musical instrument which requires three people to play it, how do people respond to it, and what is their experience of it?"

Bengler and Bryan-Kinns wanted to explore how people would respond to this novel collaborative music system, how they would engage with it, and how they might engage with each other. With these open-ended

questions, a proof-of-concept study is the most appropriate. A study was conducted over two days at the Victoria & Albert Museum with random members of the public (150+), to see how they would respond to the novel form of music making. Data was collected with many different tools, including questionnaires, observations, video recordings, and data logs from the Polymetros system. Later studies then took on a more comparative study structure, comparing how Polymetros was perceived by people in different cultural contexts - in the UK, Spain, and China (Bengler and Bryan-Kinns, 2014). Results indicate the significance of ownership and supporting individual participation in collaborative creativity; the physicality of the system's design assisted in non-verbal communication and understanding of the other players' actions and structure roles in the compositional process (Bengler and Bryan-Kinns, 2013).



(a) The layout of each Polymetros instrument.



(b) Creating collaboratively between three musicians with the Polymetros system.

Fig. 3: The Polymetros collaborative music system (reproduced from Bengler and Bryan-Kinns (2013)).

2.2 Comparative Studies

In contrast, the comparative studies we will discuss in this chapter are focused on particular aspects of interaction or effects of the MAT's use. These studies are different in the questions they ask, which are more focused on "How does X effect Y" or "Can we do...". They focus on the design of an interactive system and specific design features can change or impact users' interaction as introduced in the following examples.

2.2.1 Keppi

Keppi (Bin et al., 2018) is a Digital Musical Instrument (DMI) designed to explore the effect that disfluency in a musical instrument's design might have on performers' and audiences' perception of skill and risk in performance (Figure 4). The research question here started as a more open question of "What if I make this MAT musical instrument which is risky to play in performance as its musical properties degrade in real time" and then moved to a more focused Comparative Study question:

"What effect does increasing the disfluency in the design of a MAT musical instrument have on audience and performer perception of skill and risk in live performance?"

In this comparative study Keppi was designed and produced incorporating a disfluent design characteristic: It would turn itself off if not constantly moved. Six percussionists then performed live on stage with different versions of Keppi which each had different levels of disfluency. Audience feedback was collected in real time during the performances through an app on their mobile phones, and also through post-event questionnaires. Performer feedback was collected through survey questions, and the style of music created and performed using Keppi was analysed. The results of the study suggested that whilst different levels of disfluency in the design of Keppi did not have an effect on audience enjoyment of the performance, it did have an effect on their recognition of the skill of the performers. Moreover, performers noted that the disfluent behaviour of Keppi was viewed as a positive design feature, which contradicts conventional Human-Computer Interaction design guidelines which stress the importance of intuitive and reliable user interfaces.



Fig. 4: The Keppi, a cylindrical instrument designed with intentional disfluency characteristics (reproduced from Bin et al. (2018), with permission).

2.2.2 Daisyphone

Daisyphone is an online collaborative music editor which allows groups of people to edit a shared loop of music 48 notes long (Figure 5). The design is purposefully very simple, and the interaction is restricted to adding notes (12 pitches are possible, and 4 instrument sounds) and removing notes, and drawing to communicate (Bryan-Kinns and Hamilton, 2009). The circular area shows the shared musical loop and drawn annotations can be seen around the outside. The circular representation of the loop was chosen purposefully to be different to most music sequencers, thereby reducing familiarity with the interface. Support for shared drawing and shared editing was drawn from HCI research which stated that this would improve collaboration (in-text document editing). Bryan-Kinns and Hamilton (2009) explored the effects of having a sense of personal identity and a shared way to communicate would have on people's mutual engagement (Bryan-Kinns, 2004). Mutual engagement being "it involves engagement with both the products of an activity and with the others who are contributing to those products" (*ibid.*). The research then focused around the question of:

What role does personal identity play in collaborative musical composition, and how do different representations of personal identity compare when users complete a specific task together?

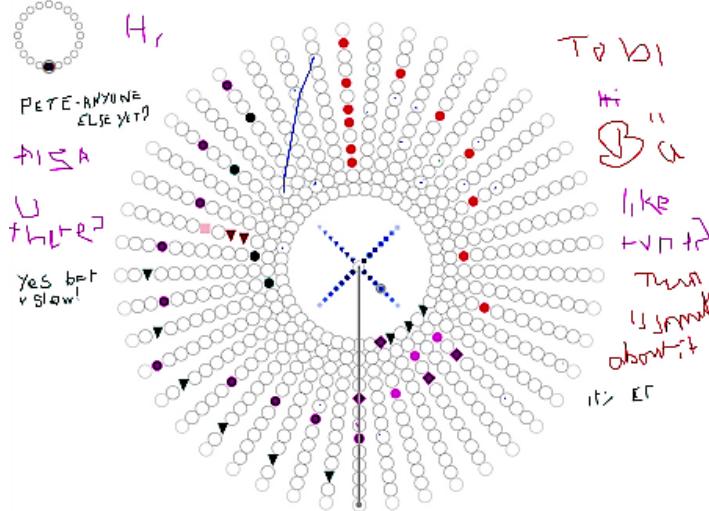
The authors undertook a Comparative Study to see what effect providing cues to personal identity in the interface and providing a shared area to communicate in would have on people's mutual engagement. The study involved 39 participants collaborating online (participants could not see or hear each other in person) to create short loops of music. Each participant spent about 1 hour in the study. Data was collected from questionnaires, and logs of interaction with both the music and the shared drawing area. Questionnaires were used to gather feedback from participants and to make comparisons between different versions of the Daisyphone interface.

2.2.3 Smart Trousers

Skach et al. (2018) designed a pair of smart trousers which were able to sense posture changes (Figure 6). This MAT involved an interdisciplinary approach combining electronics and e-textile design with behavioural and social science. The use of a wearable and the authors' relevant fashion design background allowed for the creation of a wearable that could be used to study wearers' behaviour without being disruptive to a social environment. Fabric pressure sensors were integrated into the garment and measured contact between points on the wearer's body and the legs as well as the surface of a chair. Through this garment, the authors aimed to explore different behaviour related to emotional and social communication. The study focused on exploring non-verbal communication through posture and gestures and inquired specifically as to whether gestures on the legs could be classified through pressure sensing. The research was based around a question of:

Can we use pressure sensing to gather data about the movement of the lower body when seated, and can this data be used to classify behaviours of the wearer as they engage in conversation?

This work used a Comparative Study to gather data about participants wearing the smart trousers while they performed a number of actions. In this sense, the comparison is not between one MAT and another, but rather between different users and the same wearable. The interaction with the trousers showed that, even though individual participants followed the movement directions differently, in their individual interpretation, it was possible to classify different gestures using the data gathered.



(a) The Daisypone interface on the desktop, showing player input and annotations.



(b) Compositional collaboration with Daisypone on mobile.

Fig. 5: The Daisypone interface for collaborative music-making (Fig. 5a reproduced from Bryan-Kinns and Hamilton (2009), Fig. 5b ©EPSRC, available to the public: <https://www.flickr.com/photos/epsrc/3340524049/>).

2.3 From Proof-of-Concept to Comparative Study: The Chaos Bells

Research into the interaction with large musical instruments by Mice and McPherson (2019) provides an excellent example of how exploratory proof-of-concept studies can inform further comparative research.

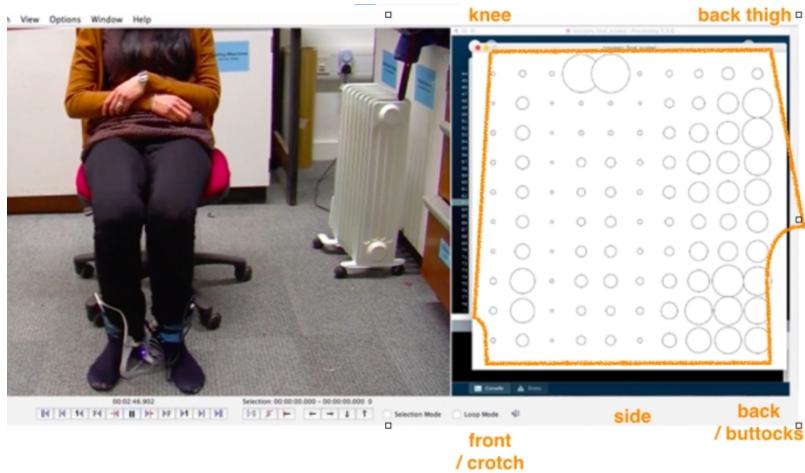


Fig. 6: Visualising the pressure measurements (right) in a participant's seated position (left) while wearing the smart trousers (reproduced from Skach et al. (2018)), with permission.

This work began with a proof-of-concept interview study which explored musicians' relationship with the physically large instruments they had already been trained on (Mice and McPherson, 2019). Interview questions focused on gestures and the precision of movements, fatigue during performance, and improvisation. In addition, participants reviewed *Cello Suite no. 1 in G Major* (J. S. Bach) and discussed difficulties they would have playing the piece transposed for their own instrument. This exploration revealed a number of insights into timbral control and the embodied relationships between the musicians and their large instruments.

From this study, the authors conducted a series of comparative studies using the Chaos Bells (Figure 7), a large-scale digital musical instrument (DMI) developed by Mice (Mice and McPherson, 2020). The design features a set of pendulums which use accelerometers which drive a Karplus-Strong algorithm as the performer strikes, raises, and swings them (Figure 7a). In further study of the instrument, the authors compare different pitch mappings on the pendulums to compare how different tonal layouts (Mice and McPherson, 2022b, 2020), instrument size, and the performer's body influence the idomatic gestures and patterns during improvisation, as well as the performers' perspectives of their own bodies (Mice and McPherson, 2022a) (Figure 7b), which show that the size of the instrument determines the gestures which can be used.

Through this work, we can see how the different study types can address different types of research questions using different approaches. The authors move between different focuses:

Q1: *How do musicians perceive their interaction with their large instruments and how might they act/feel about performing different music when playing them?* (proof-of-concept)

Q2: *Do different layouts of the tones on a large instrument impact the gestures and movements used by musicians while playing them, and does this change their perception in practice?* (comparative)

The proof-of-concept study used more open-ended approaches to gather a set of feedback and ideas surrounding the design and performance with large instruments. This was done to get a better sense of how performers work with instruments they have been trained on and know well. By using an exploratory approach, the

authors determined key factors in the interaction with such instruments. This information informed the development of a new DMI based around these principles, where more specific questions of interaction such as tonal layouts could be examined in an appropriate context.



(a) The Chaos Bells.



(b) Performer gestures working with the large instrument.

Fig. 7: The Chaos Bells, with height and width of 2 metres, designed to study the impact of interface size on interaction and gestures. Photos ©Lia Mice, used with permission.

3 Designing a MAT Study

The first step in conducting a robust MAT study is defining your research questions – Having specific questions in mind will help to structure the rest of the research into either a proof-of-concept or comparative study. As discussed in Section 2, both kinds of study are equally valid but will focus on different kinds of questions and scenarios. Having clear questions will drive consistency in the study and its analysis and focus the reporting of your results in a way which is understandable to others. This will also help to justify your design choices for both the study procedure, analysis, and the MAT — attention must be given to details of the Media and Arts themselves (e.g., musical theory, visual arts practices, traditional craft) as well as the Technology driving the MAT interaction. The media and art should be researched in detail before beginning. For instance, when working with wearable MATs, it would be important to familiarise yourself with artists' practices in fashion design, wearable technology and integration, and so on. In the presentation of both Skach et al.'s smart trousers (Skach et al., 2018, p. 117) and Reed et al.'s VoxEMG wearable implementation (Reed et al., 2022, pp. 171-172), significant consideration and direction was first drawn from previous work in textile sensing technology. The design and use of the MAT will very likely rely on background and existing work from other science backgrounds as well, in these cases fields such as cognitive and behavioural sciences, bio-mechanics, and sensorimotor interaction.

3.1 Choosing a Study Type

If your research interests are more exploratory, for instance wanting to know how people might interact with a new MAT in their typical performance practices (as done with VoxEMG and the Chaos Bells), proof-of-concept studies can provide a suitable format for your research questions. Similarly, proof-of-concept studies are good for questions such as *How do people collaborate with this MAT? How do participants' perceptions of their movement change if they use this MAT over a long time period? How do participants choose which notes to play/ which sounds to include/ which roles to take on in a duet performance?* These questions are specific, but they do not seek to understand a specific difference – rather, they are open-ended and aim to gather information about a specific context. This can be beneficial as well if you do not have any existing knowledge or want to prompt any particular behaviours. Proof-of-concept studies can be useful to observe the impact of and attitude towards a new design in-context – for instance Mazi worked with the children and their routines in their day-to-day environment (Nonnis and Bryan-Kinns, 2019b).

On the other hand, research questions that directly involve comparisons naturally warrant a comparative study. For instance, the Daisypnone study's questions revolved around a comparison of features included in an user interface. Comparative studies are more appropriate if your questions are something like *Do participants prefer one collaboration method over another? Which of these interface features is more important to experienced users, compared to novice users? or How does participants' accuracy score change after using this MAT for a month?* These research questions compare specific elements of experience and design and maybe compare responses from two or more groups of people (e.g., experts and novices). Additionally, you may be looking to gather information about different, specific moments in an interaction; for instance, with the smart trousers, examining the data gathered by the sensors at different moments in the interaction.

3.2 Designing the MAT Itself

With these research questions in mind, you can decide on what MAT you will create and then how you will study it. When designing a MAT, it is also important to consider again that the methods are rigorous and repeatable – someone else should be able to understand and recreate your design. With the growing movement of open-science, many researchers working in fields such as HCI are making their designs open-source so that they can be easily shared and accessed by other researchers and communities that will benefit from their use. Regardless of whether or not the MAT design is to be open-source, it is important to carefully document the design process to support the research in publication. Some key elements which should be included in this design documentation include:

Design Inspiration: MATs are results of inspiration-led design, often combining ideas and theories from different disciplines, and often the design focus is driven by from personal interest. For example, the collaborative music making system Daisypnone (Bryan-Kinns and Hamilton, 2009) was created as a result of a personal passion for performing music in combination with an interest in exploring new forms of networked music performance which were emerging at the time. If you are studying the design process or working in an autobiographical use case, you should note important details about your own experience and background which are relevant to the design (e.g., with VoxEMG (Reed and McPherson, 2021)).

Identify Design Features: Key design features of your MAT should be informed by previous research – your study is likely to explore a new and novel context, but it should be based in established knowledge

from the research community. For example, the design of Daisyphone was informed by Computer Supported Collaborative Work (CSCW) research, which stated that sense of identity was important in collaboration; therefore, identity was represented by colour in the user interface. This one design feature was then varied across different versions of the MAT to understand how sense of identity affected people's collaborative creativity and engagement. The research questions therefore focused on this particular aspect.

Reduce Complexity: The complexity of the interaction design should be reduced as much as possible whilst still allowing for fun, creative, and engaging interaction. This makes the MAT easier to design and build and also makes the study more focused on the novel interaction itself rather than, say, learning effects. For example, Mazi was designed to support only a very small range of musical notes making it more of a 'sound toy' than a collaborative composition system. This allowed the studies to focus on how children interacted with each other rather than focusing on, say, the composition process. Additionally, keeping the design focused on novel interaction, rather than other elements of the design, will also help to reduce confounding factors — those parameters that influence both the study conditions and the results but are not accounted for and likely not the intended parameters you want to observe. For instance, if you give participants two interfaces with different layouts to use but they involve dozens of unexplained buttons and knobs, the difficulty of the systems will confound the relationship between the instruments' layout and the participants' interaction. The results will be unclear and you will likely learn more about the participants' frustration than their preference for layout.

Decide on the Study Setting: Evaluations of MATs take place in settings ranging from controlled laboratory settings to ad-hoc settings in public. These settings give different levels of control of the study, realism of conducting the activity in-context, and appropriate data collection methods and study structures. Whilst a laboratory study gives the most control, it is the least realistic setting for creative, fun, and engaging interaction. There are trade-offs between different settings, and you should be prepared to justify the decisions you make in planning your study when discussing your research.

Focus the Activity: Even for creative activities there needs to be some focus to the interaction. Whilst evaluating MATs is not concerned with, say the efficiency or productivity of carrying out specified tasks with an interface, it is nonetheless important to focus peoples' activity in order to be able to understand their responses to the design's features. For example, with musical MATs, people need to be provided with a motivation for creating music. This might be in the form of a compositional brief; e.g., in Daisyphone studies, participants were asked to create a jingle for the Olympic Games, and in the Keppi study performers were asked to create a musical performance using the MAT.

4 Conducting and Reporting Your MAT Study

Table 1 outlines the necessary components for conducting proof-of-concept and comparative studies. We have structured this section to serve as a template to both guide the design of the research as well as its presentation (e.g., in a research paper). Following this structure and ensuring you have all of the key components in your study planning and presentation of your work will help to make the research rigorous and repeatable, and understandable by other researchers. The template sections you see here are used in the majority of research papers across scientific disciplines. This will help you to structure your presentation in a variety of venues for the topics related to your interdisciplinary work.

	Proof-of-Concept Comparative		
Study Methodology	Background	✓	✓
Research Questions	Aims	✓	✓
	Hypotheses		✓
Independent Variables		✓	
Dependent Variables		✓	
Conditions		✓	
Participants	Participants	✓	✓
Tools	Tools	✓	✓
Procedure	Procedure	✓	✓
Data Collection	Data Collection	✓	✓
Analysis	Analysis	✓	✓
Results	Results	✓	✓
Discussion	Discussion	✓	✓

Table 1: Components necessary for both types of MAT study.

4.1 Section: Background

The first step should be to conduct a literature review of existing research surrounding your topic area. The related work should create a coherent idea driving the current research. This should be done to help define research questions (e.g., whether there a gap in the existing knowledge or previous publications suggest further exploration you can address in your work) and to provide a state-of-the art for methods being used in similar studies. A strong background will help you to justify your choices for study design, methods, and data analysis, thus increasing the robustness of your work.

You should provide a brief overview of the relevant literature when presenting your research to contextualise what you are doing within an existing body of knowledge and provide rationale for your decisions. With MAT research, the background may consist of relevant work from different disciplines and so you will want to spent time to make sure that you connect literature coming from different fields. This might include defining terminology you will use, with respect to different definitions in other research, or uniting concepts from multi-disciplinary work so you can discuss them together. For instance, in Reed and McPherson (2021), the authors use the Background section of the paper to unite concepts surrounding mental imagery and embodiment from cognitive science and design practice within the context of the paper.

4.2 Section: Research Questions

As mentioned in Section 3, you should define specific research questions you wish to address through your work. With existing literature in mind, these questions will help to focus the research and the design of your MAT and study. A research question, while targeting a specific area for exploration, is usually something broad such as “How do people engage with each other when playing a three-person musical instrument?” (as in Polymetros). It is common, when writing a research article, to include the research questions and the

main contributions your work makes to the existing literature after presenting your Background. This helps to connect the Background to your current work and focus a reader on the key points you will address in the current study.

4.3 Section: Study Methodology

Referring to the above table, you should include the following components in your study design and when presenting your research. Methodology is important in order to ensure the study is robust and reproducible. The methodology should include detailed information about the procedure so that another researcher may reproduce it exactly as you did. In your presentation, you may wish to use subsections for each component of the methodology.

4.3.1 Aims

You should define and list the aims of the study. There should be a small number of aims – one or two, and definitely less than five.

For a proof-of-concept study, the aim is usually to find out people's conceptualisation or understanding of a novel piece of interaction, e.g., "The aim of the study is to explore how people engage with each other when they play a three-person musical instrument." (Polymetros).

A comparative study may examine a specific interaction paradigm, perhaps further exploring an observation from a proof-of-concept study. A comparative study example: "The aim of the study is to test whether a shared music making system with support for shared communication channels supports greater levels of mutual engagement than one without." (Daisyphone).

4.3.2 Hypotheses

If you are doing a comparative study, you will have at least one hypothesis which you test. There should be a small number of hypotheses, usually between one and five hypotheses is fine. For example, in the Daisyphone comparative study there were two hypotheses. The study examined the effect that providing a graphical annotation function as an additional communication channel has on mutual engagement:

H1: Mutual engagement would be greater where an additional channel of communication was provided – graphical annotation. (Daisyphone).

You might have a more exploratory kind of comparative study and just predict that there would be some difference (but not know what kind of difference it is):

H2: Mutual engagement would be different where an additional channel of communication was provided compared to when there is no additional channel of communication.

Make sure that, if not already done in the Background, that you define the terminology you will use here; in these hypotheses, you will want to clearly state what is meant by "mutual engagement," and how it might be measured. This should be informed by and connected to your Background section. For example, in the

Daisyphone study “mutual engagement” was described as a collaboration in which there is both “Evidence of engagement with the product of the joint activity, i.e. music in our domain. For example, participants’ reports of feeling engaged with the product, a high quality product, focused contributions, or demonstrations of skills and expertise in creating contributions.” and “Evidence of engagement with others in the activity. For example, more reports of feeling engaged with the group, coherent final joint products, colocation of contributions, mutual modification of work, discussions of quality of the joint product, repetition and reinterpretation of others’ contributions” (as described in Bryan-Kinns and Hamilton (2009)).

4.3.3 Variables

Defining variables is an important part of a comparative study plan – variables are things that are changed in your study, or things that are measured in your study (i.e., they are *variable* in the study). There are two main kinds of variables to define: *independent variables* and *dependent variables*.

You should also be aware of other elements which might act as *confounding variables*. These are things that might change between each participant in your study and which might have an effect on their performance; for instance, the skills and expertise of participants might be a confounding variable and therefore need to be controlled. If some participants are skilled musicians, their experience might have an effect on how they play a 3-person musical instrument. The time of day might be a confounding variable if people are tired in the evening versus the morning, etc. In MAT studies, it is often difficult to control these confounding variables and so they need to be included in the Discussion section of your report and you should elaborate on how they may have impacted the findings and outcomes of your study. As mentioned previously, limiting the complexity of the study will help to keep confounding variables under control.

Dependent Variables

Dependent variables are things that you measure in a comparative study. They *depend* on the interaction and what happens during the study. You should describe these as concretely as possible; for example, in conventional HCI studies, a dependent variable could be the time in seconds it takes to complete a task. For more exploratory MAT studies, your dependent variable might be more subjective; for example, in the Daisyphone study there were 9 dependent variables including: (1) “Quality measure: participants’ reports of their assessment of the quality of the final product and the collaboration itself,” (2) “Contribution to joint production measure: number of notes contributed,” and (3) “Proximal interaction measure: closeness of participants’ contributions to each other’s contributions” (as described in Bryan-Kinns and Hamilton (2009)).

For each of these, you will need to define how the dependent variable is measured or how a quality is assessed. The best way to do this is to find good definitions in existing research literature and either use those definitions or modify them to suit your studies.

Independent Variables

Independent variables are things that you change in the study; for example, changing features of the interface or the app which is provided to see their effect on user experience (as measured by your dependent variables).

You define the independent variable along with the “levels” of variable, or what things you change. For example, to test the effect of providing shared annotation in the Daisypone study, an independent variable would be “Annotation” and the 2 levels would be “Annotation” or “No Annotation” (Daisypone): “In the Annotation condition, participants could ‘draw’ on the Daisypone, and these graphical annotations were shared with other participants. In the No Annotation condition, no graphical annotation was supported, and so communication could only occur through the music.”

4.3.4 Conditions

In your comparative study you will test different variants of the interaction – each of these variants is called a *condition* or treatment in the study.

In the Daisypone study, two hypotheses were tested: **H1** “Mutual engagement would be greater where participants had explicit cues to identity” and **H2** “Mutual engagement would be greater where an additional channel of communication was provided – graphical annotation”.

Two independent variables were used to test these hypotheses: 1) cues to identity (for H1), and 2) communication channel (for H2). There were therefore 4 conditions in this study (see Table 2).

		Independent Variables	
		Communication Channel	Identity Cues
Conditions	1		
	2	X	
	3		X
	4	X	X

Table 2: The four conditions examined in the Daisypone study, showing the combination of the two independent variables. An X indicates where the variable was included in the condition.

4.3.5 Participants

Then, decide who your participants will be – are they general public, or do they need certain skills and experience? If so, what skills and experience? You will also need to decide how you will recruit the participants and if they will be paid any incentives? This information should be described in detail when you document your research. For example, the participants for the Daisypone study are described in Bryan-Kinns and Hamilton (2009):

“Final year Computer Science students at the first author’s institution were recruited through advertisements to take part in the experiment as part of their course, but not offered any incentives to take part. Thirty-nine of a possible 80 participants took part (28 males, 11 females; aged from 20 to 29 years old, mean age: 22; average computer literacy: expert; average musical ability: intermediate; none were professional or trained musicians; none had used Daisypone before). Participants’ musical preferences ranged from Hip Hop (most popular) to Latin (least popular).”

If you report individual responses in your results (e.g., interview answers), then you should give a table to provide brief demographic of each participant, as this may have an impact on their answers. Give each person a participant ID (e.g. P1, P2, etc.) which you can then use in the results. For example, see Table 3.

Participant	Age	Gender	Musical Ability	Musical Preference
P1	25	M	Novice	Hip hop
P2	24	M	Intermediate	Hip hop
P3	20	F	Intermediate	Hip hop
P4	22	M	Intermediate	Rock
P5	21	F	Expert	Latin
P6	22	M	Novice	Classical
P7	22	M	Intermediate	Rock
P8	21	F	Intermediate	Ambient

Table 3: Example table for presentation of participant demographics and background (adapted from (Bryan-Kinns and Hamilton, 2009)).

Sample Size

As a rule of thumb, you will need at least 10 participants in a proof-of-concept study. Ideally you would have much more than 10 participants for a proof-of-concept study, but it is worth noting that the most common sample size in papers in the leading HCI conference (ACM CHI) is 12 participants (Caine, 2016). For a comparative study, you need at least 10 participants for each condition, e.g. in the Daisyphone example you would need at least 40 participants as there are 4 conditions (assuming that each person only does one condition – this is called an unrepeatable-measure, also referred to as ‘between groups’).

Data suggests that, in order to discover all of the usability problems in testing, 15 participants are needed; through iterative testing, this could be done in as little as 5 participants (Nielsen and Landauer, 1993). You can reduce the number of participants needed in a comparative study by designing the study so that each participant does multiple conditions (*repeated-measures*, also called within-groups), but this makes the study longer (therefore increasing participant boredom and fatigue), increases the chance of learning effects, where participants learn something about the interaction in one condition which then either makes the other conditions easier to use or harder to use due to confusion, and potentially reduces the power of the statistical tests you can do on the data after the study.

Repeated-measures study design is beneficial in that you can ask participants comparative questions between the conditions which can help to understand how participants react differently to different conditions e.g., to compare the experience of condition 1 to their experience of condition 2.

For example, you could run the Daisyphone study with 10 participants and have each person use all four conditions. As another option, you could run the study with 20 participants and 10 of the people use conditions 1 & 2, and the other 10 participants use conditions 3 & 4. In this case, you would need to decide why it is not some other combination (e.g., 1 & 3 + 2 & 4) and again be prepared to make this justification when presenting the research. When using repeated measures, you also need to make sure to counterbalance the ordering of the conditions so that the ordering does not cause an effect in your results. You can use a balanced

Latin square to figure out different orderings for your conditions to ensure that your study is balanced¹; for instance, if you had a study with four conditions of the Daisypnone study, you could divide your measures as in Table 4.

		Trial Order			
		1	2	4	3
Participant	1	1	2	4	3
	2	2	3	1	4
	3	3	4	2	1
	4	4	1	3	2

Table 4: An example of a balanced Latin square for the four conditions of the Daisypnone study; here, the conditions are in a different order for each participant and each condition precedes another only once, ensuring that potential effects from order are removed from the study.

However, it is important to note that you may need more participants depending on what kind of data is being collected and what analyses need to be done. In order to do some statistical testing with accuracy, larger sample sizes might be needed. This should also be considered when recruiting participants and when considering how many variables are examined at one time.

4.3.6 Tools: The Media and Arts Technology Itself

If you need to build a MAT to be used in your study – for instance, an app, a VR experience, tangible interface, etc., you will need to document the design and be able to explain how it works. This would include detailing your expectation of how people would interact with it, and how it works on a technical level. You will need to connect your design to the Background to be able to define and discuss how it is similar or different to existing tools. If you are using existing software, describe what the software is and how it will be used in this specific research context (making sure to credit the original authors/developers and cite any relevant publications).

4.3.7 Procedure

With your materials and tools ready, you should then define the study procedure. Provide a step by step description of what participants in the study will be asked to do, and how long is spent on each step. Include *everything* – include the step at the beginning when you explain the structure of the study to participants (e.g. 2 minutes), include the part where you ask participants to fill in questionnaire (e.g. 5 minutes), etc. In the pursuit of repeatable studies and reproducible findings, it is important that you are detailed enough for another researcher to conduct the research exactly as you have done.

All studies should start with an introduction where you explain the purpose of the study (but don't mention what results you are looking for as this may bias their views) and that participants are free to stop at any time they like. You should also collect demographic data such as age, gender, and other relevant information such as musical experience. Specialist skills such as musical experience can be assessed using questionnaires

¹ Balanced Latin Square Generator: https://cs.uwaterloo.ca/~dmasson/tools/latin_square/

such as the Goldsmiths Musical Sophistication Index (Gold-MSI) (Müllensiefen et al., 2014) – using existing questionnaires helps to strengthen the rigor of your study. Do not collect anything that could directly identify the participant; for instance, do not collect their name, address, or email address in the demographic information. If you are studying or working in an institution such as a University you will likely need to ask participants to complete a consent form and ensure that your study has ethical clearance from your institution to proceed.

4.3.8 Data Collection

One critical component of study design is the decision on what kind of data will be needed and collected, either quantitative, qualitative, or some mixture of both. *Quantitative data* is data which is in a numerical form can be assigned a value (eg., it has a quantity — a numerical value — and is able to be measured as such). It may be participants' ratings of their enjoyment on a Likert scale from 1 to 5, or it may be more conventional HCI measures such as speed of completing an activity, or number of errors, etc. *Qualitative data* takes the form of words and is more descriptive (eg., it is more about a specific quality or descriptor of whatever is being examined). Many studies incorporate mixed methods, where both quantitative and qualitative approaches are taken. This is especially the case in your MAT research, where you might need to both collect concrete data, for instance on the MAT's computational performance or on the interaction, and also understand the emotional and aesthetic aspects of the design in a qualitative sense. It is important to consider what form your data will take when planning a study to know what kind of analyses you might use to interpret the data, and to ensure that you are able to collect enough data to get an appropriate and well-rounded analysis. As with types of MAT study, one kind of data is not a "better" option, but likely one will be better at addressing which perspectives you wish to explore and you should be prepared to justify your choices in your research presentation.

For proof-of-concept studies you will most likely collect data by writing down your observations of people's interaction with the MAT (and each other), video recording their interaction, and then having questionnaires and interviews at the end of the study to get some idea of how people responded to your MAT. For a comparative study you should describe what data you will collect for each dependent variable. Decide whether the data will be collected whilst the participant is interacting with the MAT, or after they have used it. Some examples of data collection are outlined below.

It is important to note any possible ethical considerations with your data collection and how they will be addressed. Check with your affiliated institution's General Data Protection Regulation (GDPR) or other privacy and ethics procedures. For example, if you video record people interacting with your MAT, then does your institution allow you to publish images of people's faces? If so, will you request people's consent to use their photos in papers? If they don't give consent, can you still get the data you need for your study?

Interviews

When conducting interviews with participants, consider what kinds of responses you are seeking. You might use an existing interview structure or a semi-structured interview where you decide prompts based on the topics you wish to explore. It is best to audio record interviews and then transcribe them later – people will be much more descriptive and reflective in spoken answers than in written answers.

Interviews are particularly important for proof-of-concept studies in order to get participants' feedback on novel experiences. They are also important in comparative studies to help understand why people behaved the way they did and responded to questions the way they did. You should start your interview with open questions, where you try to get participants to explain their understanding of the experience (e.g., ask: "Please could you describe what you just experienced to me?", or "Please tell me how you would describe what you just experienced to a friend?", followed by more specific question such as "What did you find most engaging about the system?", or "What did you find most challenging about the experience?").

Then, you can start to probe for more specific feedback; for example; "Please could you tell me about your experience of playing music on the three-person music instrument?", followed by "Did you find that other people responded to your musical contributions?". Make sure to follow any questions that could be answered with a yes/no answer with probing questions, (e.g., "Can you tell me why that was?" and/or "Can you give me a specific example?"). More specific and targeted questions and answers can help to explain people's responses to particular features of your MAT.

Questionnaires

It is best to use existing questionnaires, such as the Gold-MSI mentioned earlier (Müllensiefen et al., 2014), the Creativity Support Index (CSI) (Cherry and Latulipe, 2014) or the NASA Task Load Index (TLX) (Hart and Staveland, 1988). These questionnaires are validated through existing research – this means that they have been shown to be reliable in testing for certain kinds of feedback. When using such questionnaires, it is best not to select a subset of questions – use the whole questionnaire. You need to be careful to not have too many questions otherwise participants will become bored or frustrated, especially in online studies.

Make sure to choose your questions and questionnaires carefully to address your research questions. You will need to balance your time: not too long for participants to complete without losing focus, and not too short that you don't get any useful data. As a rule of thumb, you are likely to need at least 10 questionnaire questions to get useful data. Try to restrict the length of questions to one A4 side if possible. In the Polymetros example a 7-question questionnaire was printed on one side of A5 paper, but this was due to very short time for people to complete the questionnaire in a high traffic public venue and resulted in very little useful data.

Video-Cued Recall

You might also ask participants to watch a video recording of their interaction experience and provide a commentary on it. This provides some reflective data on what they did and how they responded to the interaction. To provide results which are more comparable between participants, you might focus on getting participants to identify particular pieces of behaviour you are interested in, (e.g., points at which they learnt a new aspect of the interaction, felt frustrated, introduced new ideas, or they felt most immersed in the experience).

Observations and Retrospective Video Analysis

Observations are usually carried during the interaction to get an overview of the forms of interaction with the MAT. You would then go over video recordings of the interaction to annotate the video with descriptions of

the interaction and then code the interaction. For example, you could analyse video in terms of the following coding schemes depending on your study, or develop your own coding scheme:

- Participants mirroring, transforming, or complementing each other's contributions or actions. This is referred to as evidence of Mutual Engagement (Bryan-Kinns and Hamilton, 2009).
- Changes in participants performative interaction - whether they are simply observing, or participating, or performing (Sheridan and Bryan-Kinns, 2009).
- Number of new ideas generated e.g., ideas in a brainstorming session.
- Topics of conversation between participants - what are the conversations predominantly about? The system, the creative act, each other, the organisation of the activity, the weather? If there is a lot of talking between participants, you could use Thematic Analysis (see Section 4.4) to identify the common themes.

4.4 Section: Data Analysis

Data analysis will generally be broken into either quantitative or qualitative analysis. Generally, decide and describe what kinds of data analysis you plan to do with the data collected. Often, this will not be perfectly clear before you get the results, but you should provide an indication of the kinds of data analysis you plan to do so that you can collect the right kind of data for the kinds of analysis you want to be able to do. This will be largely dependent on what you want to know from your study.

Data Measure Types

When collecting data from participants through questionnaires or other feedback mechanisms, you need to specify what kind of measure it is. You will come across the following terms:

Nominal data, or categorical data, uses labels for variables without giving a quantitative value; for instance, having participants indicate gender or using arbitrary categories (e.g., Interface A or Interface B). This data is separated in distinct categories which cannot be ranked.

Ordinal data is similar, but the categories follow a natural order; for instance, asking participants the highest education level they have achieved. There is an order from compulsory schooling to undergraduate to postgraduate studies. This could also include ranking where things labeled as first, second, third preference.

Interval data, or integer data, is measured quantitatively but there is no zero point and it can be negative. The difference between the values on the scale should be measurable and comparable. Age is a common interval data value.

Ratio data is similar but there is a zero-point restricting the range. For instance, asking a participant how many times they use a software during the course of a day. Time duration is also a common ratio data, for example, measuring how long a participant took to complete an activity.

Statistical Testing

For quantitative data, you should decide the statistical tests to perform on the data to see if your observed results match your hypothesis. We use statistical tests to determine whether there are significant relationships between variables or difference between groups. By *significant* here we mean that the differences are not likely to be just due to chance. Ensure that you have enough participants with the variables you are examining, as mentioned in Section 4.3.5.

You can use *Regression Testing* to check potential cause-and-effect relationships, comparison tests to check for differences or similarities in groups and their behaviour, or correlation analyses to see whether variables are related. When performing statistical analysis, make sure that your data fits into the assumptions of the test being used (if you use a parametric test). Transforming your data can help to achieve normal distribution and variance. A common comparative test used in MAT work is an *Analysis of Variance* (ANOVA); this kind of test examines the difference in two or more groups based on the mean of their data.

Thematic Analysis

With qualitative data, you will need to decide on an evaluation method for the interviews and open-ended questionnaires. Interviews should be transcribed so that you can anonymise and analyse the text data. In MAT research, we commonly use Thematic Analysis on interviews and open-ended questionnaires to identify key themes in people's responses to MATs (Braun and Clarke, 2006, 2012). This method is particularly useful for qualitative research as it produces results which are useful for further understanding, rather than qualitative results such as frequency of responses (Braun and Clarke, 2020). For instance, a single participant might make an important comment about their interaction which is different from the others' perspectives and this would be included as an important part of the analysis, rather than as an outlier. You will also need to consider how you will analyse people's behaviour, such as gestures and movement; for instance, coding movement in Laban notation (Laban and Ullmann, 1971; Loke et al., 2005).

Analyses should be presented referencing the specific method used. This Methods section should only include how the analyses were conducted, saving the results until the subsequent Results section. Here, it is important to describe different measures taken from the data; for instance, how participant involvement was measured, how musical complexity was determined, and so on.

Triangulation

Additionally, the Methods section is an ideal place to discuss why different analysis methods were chosen; for instance, justifying the choice of a statistical test. If you want to know about more users' qualitative perceptions of their interaction, it is probably not necessary to conduct statistics testing (and you may not have data which you could test in this way). Therefore, analysing participant responses is more appropriate. If you want to know about more quantitative differences between one interface or another, you could use statistical tests; for instance, seeing if there are more annotations made on one kind of interface than the other (as with Daisypnone).

In most studies, you will want to have a bit of both kinds of data – statistical results are used to support the qualitative observations made, and qualitative results are used to contextualise the numerical data gathered. This practice is referred to as *triangulation* — triangulation is an analysis practice which uses multiple

analysis techniques and data streams to get a clear, robust, and well-rounded picture of the study². In the DaisypHONE example, quantitative information was collected about the participant's annotation behaviour (number of annotations, locations marked, timing of annotation, etc.) through the program itself. Interviews were also used to gather open-ended response data about the participants' experiences with DaisypHONE. Together, the triangulation of these results tell us not only what was going on, but also why. For instance, qualitative analyses of survey responses after working with Polymetros, which suggest controllability and appropriate interest by broader participant groups, was highly associated with participants' responses of feelings of being in control and enjoying the collaborative creative process. (Bengler and Bryan-Kinns, 2013, pp. 238-239).

4.5 Section: Results

Results should be presented based on the types of analyses done. If you performed multiple analyses, for instance a statistical analysis of study outcomes and then a Thematic Analysis of participant feedback, the results section should be divided into separate parts. This can also help to explain the triangulation of the analysis, making sure to explain each approach and analysis separately before connecting them in the subsequent Discussion section.

4.5.1 Reporting Statistical Tests

Depending on which format your publication is in, the presentation of statistical results will be slightly different, but they generally follow the same formats (make sure you check the formatting for the style of your presentation). For instance, with an ANOVA we would report something like:

$F(\text{between group DoF}^3, \text{within group DoF}) = \text{the F statistic}, p = \text{p-value}$

Example: $F(2, 26) = 8.76, p = .012$.

The p-value

The p-value is the probability that the result of the statistics test was due to chance. In HCI, we usually consider something to be statistically significant if the p is less than .05 ($p < .05$)⁴. You will also see other p-values of $< .01$ and $< .001$ – these are not typically used in HCI research (see below for more information on the confidence interval).

It is worthwhile to mention that there is some debate about p-values being viewed as the end-all-be-all in results reporting. When you examine the results of your statistics tests, it is important to use critical thinking about your analyses and interpret the p-value in an appropriate way. As mentioned before, your research will be much stronger if you can connect the results of your statistics testing to other data you collected – in a

² Note that, although it is called *triangulation*, this does not mean that specifically three methods must be used

³ Degrees of freedom, the maximum number of independent values than can logically occur in your dataset

⁴ NB: In APA formatting, you should not include a leading 0 when reporting p-values because they exist only between 0 and 1.

way, the statistics results are meant to support the validity of the observations you made in your study (e.g., in the Polymetros example of triangulation (Bengler and Bryan-Kinns, 2013)).

4.5.2 Questionnaire Results

Generally, when reporting questionnaire results, you want to report either a data count or a range-mean-standard deviation set. What you report depends on the kind of data. For nominal and ordinal data, you will generally want to report a count; for example, “All participants indicated the interface was Enjoyable (18 participants) or Very Enjoyable (30 participants) to use.” For interval and ratio data, you will want to include the range of the response, the mean, and the standard deviation; for instance, “Participants reported that they used the new interface for a longer duration of time, ranging between 45-65 minutes ($M = 50$, $SD = 2.5$).”

You can visually report the results of Likert scale questions using a box-and-whisker plot. These are good at showing many aspects of your data – they show the minimum, the maximum, the median, the upper quartile (top 75%) and lower quartile (lower 25%) of your results. For example, the box-and-whisker plot in Figure 8, referenced from the Polymetros study, shows responses given for elements of playing experience (Bengler and Bryan-Kinns, 2013).

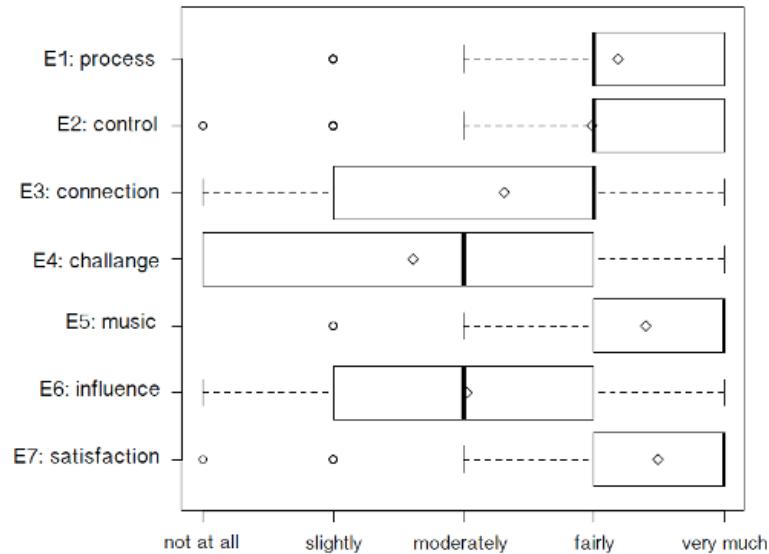


Figure 5. Items related to playing experience (150 participants)

Fig. 8: Presentating results using a box-and-whisker plot (reproduced from Bengler and Bryan-Kinns (2013)).

Participant Quotes

It is often helpful to provide quotes from participants to illustrate your results, especially when reporting the results of questionnaires and interviews. This provides evidence that you have objectively collected opinions about your MAT. In addition, referencing specific data points and codes when describing a theme from a thematic analysis can help to better contextualise the material and the meaning of the theme. When referencing specific points, make sure to include the participant's ID number.

Quotes should be short and concise to support specific points, only choosing the relevant portion of the feedback, e.g., if a participant said or wrote in their response: "Well, I enjoyed the experience a lot and found it quite engaging. I was thinking about it the other day whilst walking to work and thought that it was quite nice. That is just my opinion of course, but in general that is my feeling about it. Yes." (P2), you should not quote the whole paragraph, but instead quote the key points, e.g., "I enjoyed the experience a lot and found it quite engaging" (P2). The other points are just repeating this point. If you are really short of space, you can reduce it further, e.g., "enjoyed the experience a lot... quite engaging" (P2).

Make sure to include a range of participants' responses – don't just report one or two participants. If you have 5 participants, make sure to provide quotes from each of them in your results. If you have more than 5 participants then try to report a balance of responses – positive as well as negative. Remember that negative responses are still useful in terms of research, and give you something interesting to discuss in your discussion section e.g., to discuss why you think that they gave negative responses etc.

Codes from Thematic Analyses

When reporting a qualitative method such a Thematic Analysis (TA), it is useful to include the level of detail which was transcribed and analysed (e.g., did you include non-verbal utterances such as laughter, facial expressions, etc.).

Report how many codes you developed in the analysis and provide a general overview of the themes (you do not need to list the codes, but this may be helpful in getting a clearer view of the data). List the themes that you identified in the TA. You can also list the number of codes within each theme to give an idea of how many times the theme was found in participant responses:

E.g., "The 949 coded segments were clustered into the codes: effort; entanglement; characteristics of the compositions; reflections on the instrument; gestures and techniques; performing perception; performer's body; movement; learning the instrument over time; and 'edge-like interactions'" (Mice and McPherson, 2022a, p. 5).

Then provide a description of each of the themes. The description is usually at least a couple of sentences, and if you have space it could be several paragraphs. In the description make sure to illustrate your description with quotes from participants and refer to the participant ID. Remember that TA is about providing a cohesive picture of the data from the codes, so the themes should be linked together and discussed thoroughly. For instance, the *performer's body* theme from Mice & McPherson is further elaborated as:

E.g., "During the course of the study, we noticed examples of participants feeling differently about their bodies while performing the instrument. Some comments were overwhelmingly positive, for example P5 said the instrument makes her body feel powerful, while other comments implied that participants would like to change their bodies to be more suitable for performing such an oversized

instrument. P5 said performing the instrument “makes me want more arms”, and P8 commented “I need bigger arms”. P9 said “I wish I had 3 hands.” (Mice and McPherson, 2022a, p. 12).

Remember that the number of codes within a theme does not dictate relevance; the themes are meant to create a clear picture of the data and capture important points and similarities. If the theme contains only one or two codes but provides a critical observation from the research, it is still a valid theme (Braun and Clarke, 2020). With subjective research, participant responses when working with MATs will often demonstrate a wide diversity of interaction perspectives and techniques, and it is important to give attention to this variability; for instance, participant responses when working with the Keppi (Bin et al., 2018, pp. 49-50) or Chaos Bells (Mice and McPherson, 2022a, pp. 9-12) are discussed in detail to represent the varying viewpoints when working with the instruments.

4.5.3 Other Observations

It may be beneficial to provide other observations made during the study which do not fall into a specific category of analysis. For example, providing vignettes of observations that support or illustrate findings from your other data. Vignettes are short descriptions of some action and interaction and would often be accompanied by an image from your video recording if you have it.

E.g., from (Bengler and Bryan-Kinns, 2013): “A prevalent input strategy was the creation of musical patterns characterised by simple geometric properties. The most common phrases consisted of horizontal and upward or downward diagonal lines whereas in most cases all available notes were used (Figure 9). Resulting in ‘closed musical figures,’ this approach was applied by many players providing a clear audio-visual correlation between the representation on the interface and the musical result.”

4.6 Section: Discussion

In the Discussion, you should reflect on your results – what results were unexpected, why do you think that might be? What explanations can you think of for the results you found? Connect your results back to the Background and the aims of your study – did your results match with the results of other researchers you mentioned in the literature review? If not, why not? What do your results tell you about the things you wanted to find out in the aims of your study?

It is quite usual to find things that you did not expect, or which are surprising and counter intuitive. This is part of why we conduct these exploratory MAT studies. The Discussion can explore why the results were unexpected and suggest future work to further explore the findings in future work (hopefully using your well-reported and rigorous methods). The most important thing is to clearly and objectively report the results, and then reflect on your results, connecting them to your literature review, your study aims, and your intuition about what happened.

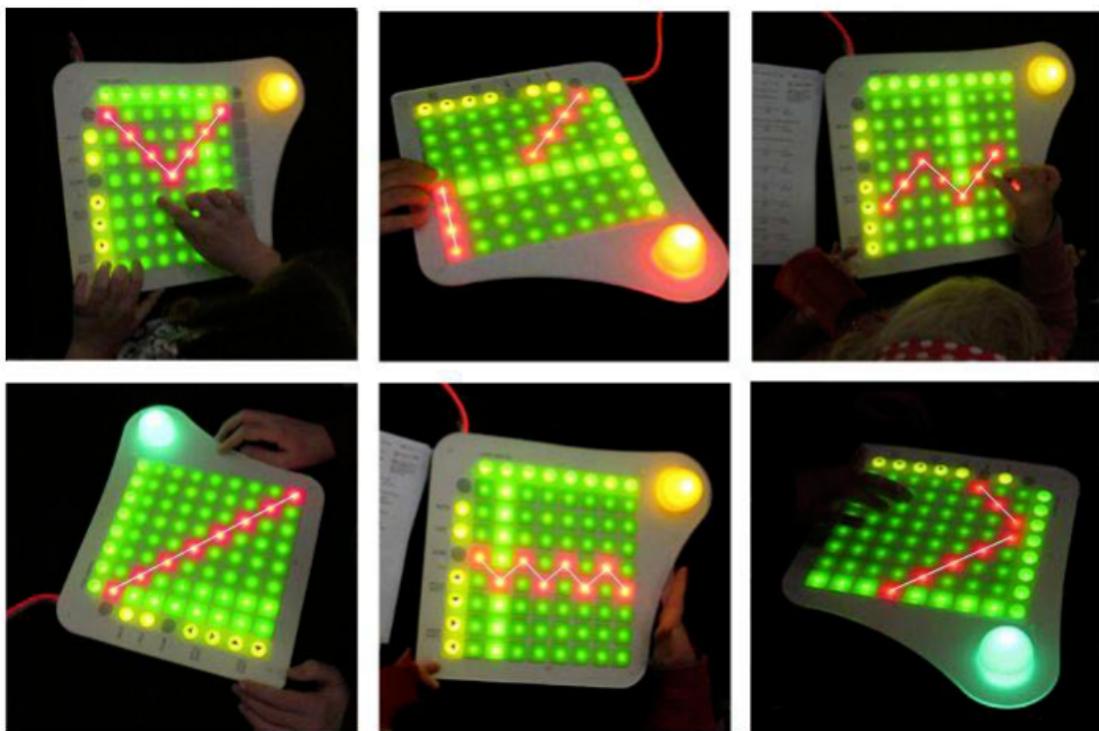


Fig. 9: A vignette of typical contributions with Polymetros (reproduced from Bengler and Bryan-Kinns (2013)).

4.6.1 Limitations

The Discussion is also the place to discuss limitations to your research. In terms of *internal* and *external validity*. Ideally, your research will have both; in the case it does not, you will need to highlight where the reported results are limited.

Internal validity is the extent to which the study accurately and confidently depicts the relationship between the variables being examined. Some factors which might influence the internal validity are uncontrolled or confounding variables, small sample size, repeated testing and learning effects and potential experimenter bias.

External validity refers to the extent that the findings of the study can be applied in other settings – how generalisable the findings are to the world outside of the study. Some factors which might reduce the external validity are selection bias (e.g., the participants who take part in a MAT study might be interested in or already using technology in artistic settings, so might react differently than others who are not as likely to participate), situational factors such as time of day and location, and limited examined factors, such as only looking at music within the Western canon.

In either case, you must identify these factors in validity, not only to lead future work but also to demonstrate your awareness and acknowledgement of potential limitations in the work. For instance, in Reed and McPhereson (2021, pp. 8-9), the autoethnographic approach means that, while the results are useful for design and interaction research, the specific interactions discussed in the paper may not be generalisable to other singers with their individual perspectives:

“It is of course critical to again state that, while the interaction observed provides a detailed account of a prolonged interaction with biofeedback through sEMG, this interaction is highly specific to the user... As mentioned previously, further studies to conduct similar trials and autobiographical use cases with the system will be necessary to validate the universality or differences in the experiences.”

4.6.2 Future Work

After providing the results and having discussed some limitations, it is worthwhile to suggest future research studies which will further explore the results you have achieved or address the limitations. You might also include some suggestions for research based on interesting or exciting findings, to expand or apply what was learned through the study to additional research questions.

4.7 Other Presentation Components

When presenting your research, you will likely need to include an Abstract, Introduction, and Conclusion to bookend the components outlined here. These portions of a paper are sometimes written last, after all of the other information is in place, to better summarise everything in the presentation together.

After the main body of the paper, you may also wish to present your materials alongside your research and results in an appendix. This will help with reproducibility and allow others to follow your methods in their own work. In your appendix you could therefore include all questionnaires, exactly as given to participants and lists of interview questions.

5 Conclusion

Research in Media and Arts Technology (MAT) must strike a careful balance between arts and science practices and norms. This can be a difficult and yet rewarding balance to achieve. However, the two are not mutually exclusive and the interaction between HCI, User Experience, and Interactive Arts and Media have led to the advancement of both fields. While scientific approaches make the research grounded and extendable to other fields, arts approaches offer new inspiration, contextualisation, and opportunities to explore many facets of the human condition. Through this guide, we have introduced you to scientific practices to make the design of MATs, their study and evaluation, and their presentation and dissemination to others both rigorous and repeatable. Through this guided approach, we hope to increase the accessibility and validity of MAT research in wider artistic and scientific communities and to further explore the human condition of being in the world.

Acknowledgments

Work on this chapter was supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1). We would like to thank all the students and staff at the Media and Arts Technology Centre and the Centre for Digital Music at Queen Mary University of London who have contributed to ongoing discourse about the nature of Media and Arts Technology which inspired many of the thoughts in this chapter.

References

- Benford, S., Greenhalgh, C., Giannachi, G., Walker, B., Marshall, J., and Rodden, T. (2013). Uncomfortable user experience. *Communications of the ACM*, 56(9):66–73.
- Bengler, B. and Bryan-Kinns, N. (2013). Designing collaborative musical experiences for broad audiences. In *Proceedings of the 9th ACM Conference on Creativity & Cognition (C&C'13)*, pages 234–242. ACM.
- Bengler, B. and Bryan-Kinns, N. (2014). In the Wild: Evaluating Collaborative Interactive Musical Experiences in Public Settings. In *Interactive Experience in the Digital Age*, pages 169–186. Springer International Publishing.
- Bin, S. M. A., Bryan-Kinns, N., and McPherson, A. P. (2018). Risky business: Disfluency as a design strategy. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 45–50, Blacksburg, Virginia, USA.
- Blythe, M. A., Overbeeke, K., Monk, A. F., and Wright, P. C., editors (2004). *Funology: From Usability to Enjoyment*. Springer.
- Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.
- Braun, V. and Clarke, V. (2012). Thematic Analysis. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D., and Sher, K. J., editors, *PA Handbook of Research Methods in Psychology*, volume 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological. American Psychological Association, Washington.
- Braun, V. and Clarke, V. (2020). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3):328–352.
- Bryan-Kinns, N. (2004). Daisyphone: The Design and Impact of a Novel Environment for Remote Group Music Improvisation. In *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '04, page 135–144, New York, NY, USA. Association for Computing Machinery.
- Bryan-Kinns, N. and Hamilton, F. (2009). Identifying Mutual Engagement. *Behaviour & Information Technology*, 31(2):101–125.
- Caine, K. (2016). Local Standards for Sample Size at CHI. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, pages 981–992.
- Candy, L. (2011). Research and Creative Practice. In Edmonds, E. A. and Candy, L., editors, *Interacting: Art, Research and the Creative Practitioner*, pages 33–59.
- Candy, L. and Edmonds, E. (2018). Practice-Based Research in the Creative Arts: Foundations and Futures from the Front Line. *Leonardo*, 51(1):63–69.
- Candy, L. and Ferguson, S., editors (2014). *Interactive Experience in the Digital Age*. Springer.

- Cherry, E. and Latulipe, C. (2014). Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Trans. Comput.-Hum. Interact.*, 21(4).
- Duarte, E. F., Merkle, L. E., and Baranauskas, M. C. C. (2019). The Interface between Interactive Art and Human-Computer Interaction: Exploring Dialogue Genres and Evaluative Practices. *Journal of Interactive Systems*, 10:20.
- England, D. (2016). Art.CHI: Curating the Digital. In *Curating the Digital*, pages 1–7. Springer International Publishing.
- Hart, S. G. and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, pages 139–183. Elsevier.
- Höök, K., Caramiaux, B., Erkut, C., Forlizzi, J., Hajinejad, N., Haller, M., Hummels, C. C. M., Isbister, K., Jonsson, M., Khut, G., Loke, L., Lottridge, D., Marti, P., Melcer, E., Müller, F. F., Petersen, M. G., Schiphorst, T., Segura, E. M., Ståhl, A., Svanæs, D., Tholander, J., and Tobiasson, H. (2015). Embracing First-Person Perspectives in Soma-Based Design. *Informatics*, 5(8):1–26.
- Jeon, M., Fiebrink, R., Edmonds, E. A., and Herath, D. (2019). From rituals to magic: Interactive art and HCI of the past, present, and future. *International Journal of Human-Computer Studies*, 131:108–119.
- Laban, R. and Ullmann, L. (1971). The mastery of movement.
- Loke, L., Larssen, A. T., and Robertson, T. (2005). Labanotation for Design of Movement-Based Interaction. In *Proceedings of the Second Australasian Conference on Interactive Entertainment*, IE '05, page 113–120, Sydney, AUS. Creativity & Cognition Studios Press.
- Mice, L. and McPherson, A. (2019). Embodied Cognition in Performers of Large Acoustic Instruments as a Method of Designing Large Digital Musical Instruments. In *Proceedings of Computer Music Multidisciplinary Research Conference, Marseille, France*, Marseille, France.
- Mice, L. and McPherson, A. (2020). From miming to NIMEing: the development of idiomatic gestural language on large scale DMIs. In Michon, R. and Schroeder, F., editors, *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 570–575, Birmingham, UK. Birmingham City University.
- Mice, L. and McPherson, A. (2022a). Super Size Me: Interface Size, Identity and Embodiment in Digital Musical Instrument Design. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–15, New York, NY, USA. Association for Computing Machinery.
- Mice, L. and McPherson, A. (2022b). The M in NIME: Motivic analysis and the case for a musicology of NIME performances. In *Proceedings of International Conference on New Interfaces for Musical Expression, Waipapa Taumata Rau, Auckland, Aotearoa New Zealand*.
- Müllensiefen, D., Gingras, B., and Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE*, 9(2):e89642.
- Nam, H. Y. and Nitsche, M. (2013). Interactive installations as performance. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction - TEI '14*. ACM Press.
- Neustaedter, C. and Sengers, P. (2006). Autobiographical design in HCI research: designing and learning through use-it-yourself. In *Proc. DIS 2012, June 11-15, 2012, Newcastle, UK*, pages 514–523.
- Nielsen, J. and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of ACM INTERCHI'93 Conference, Amsterdam, The Netherlands, 24-29 April*, pp. 206–213.
- Nonnis, A. and Bryan-Kinns, N. (2019a). Mazi: A Tangible Toy for Collaborative Play between Children with Autism. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, IDC '19, page 672–675, New York, NY, USA. Association for Computing Machinery.
- Nonnis, A. and Bryan-Kinns, N. (2019b). Mazi: Tangible Technologies as a Channel for Collaborative Play. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

- Reed, C. and McPherson, A. (2020). Surface Electromyography for Direct Vocal Control. In Michon, R. and Schroeder, F., editors, *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 458–463, Birmingham, UK. Birmingham City University.
- Reed, C. N. and McPherson, A. P. (2021). Surface Electromyography for Sensing Performance Intention and Musical Imagery in Vocalists. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '21, New York, NY, USA. Association for Computing Machinery.
- Reed, C. N., Skach, S., Strohmeier, P., and McPherson, A. P. (2022). Singing Knit: Soft Knit Biosensing for Augmenting Vocal Performances. In *Augmented Humans 2022*, AHs 2022, page 170–183, New York, NY, USA. Association for Computing Machinery.
- Sheridan, J. G. and Bryan-Kinns, N. (2009). Designing for performative tangible interaction. *International Journal of Arts and Technology*, 1(3-4).
- Skach, S., Stewart, R., and Healey, P. G. T. (2018). Smart Arse: Posture Classification with Textile Sensors in Trousers. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 116–124, New York, NY, USA. Association for Computing Machinery.