Sile O'Modhrain
Sonic Arts Research Centre
Queen's University Belfast
University Road
Belfast, BT7 1NN
Northern Ireland
sile@qub.ac.uk

A Framework for the Evaluation of Digital Musical Instruments

At the outset of a discussion of evaluating digital musical instruments (DMIs)—that is to say, instruments whose sound generators are digital and separable (though not necessarily separate) from their control interfaces (Malloch et al. 2006)—it is reasonable to ask what the term *evaluation* in this context really means. After all, there may be many perspectives from which to view the effectiveness of the instruments we build. For most performers, performance on an instrument becomes a means of evaluating how well it functions in the context of live music making, and their measure of success is the response of the audience to their performance. Audiences evaluate performances on the basis of how engaged they feel by what they have seen and heard. When questioned, they are likely to describe good performances as "exciting," "skillful," "musical." Bad performances are "boring," and those which are marred by technical malfunction are often dismissed out of hand.

If performance is considered to be a valid means of evaluating a musical instrument, then it follows that, for the field of DMI design, a much broader definition of the term "evaluation" than that typically used in human–computer interaction (HCI) is required to reflect the fact that there are a number of stakeholders involved in the design and evaluation of DMIs. In addition to players and audiences, there are also composers, instrument builders, component manufacturers, and perhaps even customers. And each of these stakeholders may have a different concept of what is meant by "evaluation." Composers, for example, may evaluate an instrument in terms of how reliable it is. If a composer writes a piece of instrumental music to be performed on a DMI, then they ought to be able to assume that (1) the instrumentalist is skilled on their instrument, and (2) the instrument has a known space of sound attributes that the composer can draw upon for musical effect.

Computer Music Journal, 35:1, pp. 28–42, Spring 2011

The designer of a DMI, who may also be a composer and/or performer, is primarily interested in ensuring that the instrument does what it was intended to do—in other words, if the instrument is designed to respond to certain gestures of the player, that it does so in a reliable way. However, a designer may also wish to leave room in their design for a skilled player to explore the "corners" of an instrument's sound space, much as a skilled violinist can exploit extended playing technique that expands the range of bowing and fingering gestures.

For manufacturers of DMIs or components of DMIs, evaluation means testing the reliability of the systems they build at a much lower level. Their motivation is primarily financial because they must determine whether the system or any component of the system is likely to fail and cost money to repair or replace. Customers, too, engage in a form of evaluation by voting with their wallets. If the product has flaws in its hardware design, its interaction design, or the quality of the sound it produces, then it simply will not sell.

These examples suggest that DMI designs can be evaluated from multiple perspectives, each of which may require different techniques and approaches. Furthermore, boundaries between roles, although usually distinct for acoustic instrument development, are blurred in the world of DMI design. Performers are often the composers of the music they play and may also be the designers of their instruments. This poses additional evaluation challenges because it requires the digital instrument builder to identify which role they must take on in objectively critiquing their work. Given that there is no one-size-fits-all solution to evaluating DMIs, a next step in determining what approaches are appropriate for a given context is to ask what such evaluations seek to discover and why. In reviewing existing examples of evaluations of DMIs, it quickly becomes apparent that answering this question has given rise to a variety of methodological approaches to evaluation. It is therefore important in each case to bear in mind that the results obtained will

reflect not just the question posed, but also the methodological approach used and the interests of the stakeholder carrying out the evaluation. The framework proposed in this article is an attempt to provide a structure within which these competing interests can be reflected, enabling performers, designers, and manufacturers to more readily identify the goal of a given evaluation, and view their own methodological approach within the context of prior work.

Disentangling Some Terminology

Before embarking on a discussion of evaluation, it may be helpful to define some terms that are often used in describing methodologies for designing and evaluating systems in HCI, as these terms also frequently appear in the literature on DMI design and evaluation.

Guidelines and Principles

Much research in HCI culminates in lists of guidelines and/or principles for design and/or evaluation of design. Principles represent general statements that are based on research or practical experience relating to how people learn and work (Dumas and Reddish 1999). They represent general design goals, but provide little guidance about how such goals should be achieved. Cook's "principles for designing computer music controllers," for example, include statements such as: "Instant music, subtlety later," a statement that says nothing about how to achieve this goal in an instrument design, just that such a goal is a desirable property of a successful computer music controller (Cook 2001, 2009).

Guidelines, on the other hand, are often derived from general principles and define practical goals that can be applied to particular contexts of use. Wanderley and Orio (2002), for example, provide a set of guidelines to aid in selecting suitable tasks for evaluating DMI designs. These are more specific than principles and are intended for a specific context—DMI evaluation. Although guidelines and principles are not in and of themselves methods for evaluating DMI designs, they do represent abstractions of evaluation results or observations from practical experience that can be used to guide future work.

Models

Models are representations of systems or artifacts that provide a means of reflecting upon the design or behavior of a system. Models can take many forms, from mathematical formulae that describe the structure of galaxies, to flow diagrams that describe system behavior, to miniature physical models of buildings and cities that make explicit topographic features and spatial relationships. They exist along a continuum ranging from descriptive models that employ analogy and metaphor at one end to predictive models that employ mathematical equations at the other, with most models lying somewhere in between (MacKenzie 2003). A particular class of models of interest here are cognitive or mental models, namely, the models constructed by users that enable them to understand how physical systems work. A mental model is defined as an internal representation of an external reality. It is built on-the-fly, from knowledge of prior experience, information acquired via perception, and problem-solving strategies. Such a model contains minimal information, is unstable, and is subject to change. Its purpose is to guide decision-making in novel situations and to provide feedback on such decisions. Mental models also allow users to rehearse actions and predict potential outcomes of these actions (Gentner and Stephens 1983).

In the context of HCI, authors such as Norman (1988) and Cooper (1995) have proposed that, in developing a novel computer system, three coexisting representations of the system can be usefully modeled in order that the interaction between these representations can itself be fully understood.

1. The system or implementation model—the model that describes the workings of the system from the perspective of the hardware and/or software designer.

2. The user's model—a description of how the user understands the behavior of the system, e.g., how they model the relationships between cause and effect, between the actions they perform and the system's response.
3. The design model—the way in which the designer represents the behavior of the system to the user including the presentation of possibilities for action, the behavior of the system throughout an interaction, and the representation of relationships between different system components.

In the context of DMI design, models have been employed for many purposes. The synthesis technique known as *physical modeling* (Smith 2008), for example, has been used to predict the behavior of every aspect of an instrument, from the resonance of the instrument's body to the behavior of its excitation mechanism, e.g., a bowed string (Serafin 2004). Other aspects of DMI design have also benefited from the development of models. In particular, the knotty problem of mapping the gestures of a performer onto the parameters that control the sound output of an instrument has been addressed through a series of modeling papers that seek to constrain the space of mappings to those that are useful for the instrument designer. These models range from strategies for directly mapping input or control parameters into parameters for sound synthesis (Choi, Bargar, and Goudeseune 1995), to more complex models that incorporate intermediate layers of abstraction (Hunt and Kirk 2000). A further reason to employ models as part of the process of defining mappings in DMI is highlighted by Fyans, Gurevich, and Stapleton (2009). Motivated by a desire to model a spectator's understanding of error in performance on DMIs, they suggest that the presence of a clear design model that relates a performer's gestures to an instrument's response will determine the degree to which the spectator can build their own mental model of the interaction and thereby understand the performer's intent.

Frameworks

The term *framework* is often used in the context of HCI to describe a conceptual scaffold that can help to elucidate relationships between design approaches within a given design space. Frameworks that are generative in nature, that is, those that lay out spaces of design possibilities, serve to systematize thinking and promote reflection that may uncover new design ideas. In defining his interaction design framework, for example, Verplank (2003) creates a scaffold to guide those tasked with developing a design idea by setting out the relationships between different aspects of an interaction and the overall conceptual model governing the design. Verplank stresses that "the invention of an interaction involves not only one compelling scenario and a unifying metaphor but consideration of a variety of scenarios and a wide exploration of alternative and mixed metaphors" (Verplank 2003, p. 8).

Other frameworks have been developed for HCI that are more evaluative in nature, typically incorporating guidelines and principles, and binding them together within an overarching design specification. Hornecker and Buur (2006), for example, propose a framework for tangible interaction design that makes explicit the relationship between four emergent "themes" in the field of tangible and embedded interaction. The authors explicitly state that their purpose is not to provide a taxonomy for designing tangible interfaces, but to posit perspectives and themes for the analysis and conceptual guidance for design.

Frameworks need not only be expressed in terms of structures that organize previously defined thematic areas, design guidelines, or design principles. Bellotti et al. (2002), for example, develop a framework for designing sensing systems that is structured as a set of questions posed to the designer, each of which is then broken down into a number of challenges that should be considered. As they point out, designers of interactions that rely on sensing systems are presented with questions that differ somewhat from those of standard window, icon, menu, pointing device paradigms, e.g., "When I address a system, how does it know I am addressing it?" In the graphical user interface paradigm, this

is a relatively trivial problem—the user moves a pointer onto an interface object, clicks on it, and the system responds. For a sensor-based system such as a DMI, this design challenge expands into a whole series of sub-challenges such as:

1. How does the system disambiguate signal from noise?
2. How does the system identify the intended target for an action?
3. How does the system know what to ignore? (Adapted from Bellotti et al. 2002).

By constructing a framework that addresses such questions in a systematic way, these authors provide a useful tool against which the performance of a sensing system can be tested.

Frameworks for DMI design have been proposed that address domain-specific issues, such as the mapping of physical input controls to sound synthesis parameters (Gelineck and Serafin 2009), and for contextualizing paradigms for DMI performance and control (Malloch et al. 2006). The framework presented by Malloch and colleagues builds upon Rasmussen's model of human information processing, in which both performance behaviors and performance context are characterized as belonging to model/symbol, rule/sign, or skill/signal domains (Rasmussen 1986, cited in Malloch et al. 2006). In the context of their framework, Malloch et al. suggest that skill or signal mode is the mode most similar to what is normally understood as musical interaction because it is characterized by rapid, coordinated movements in response to continuous sound or haptic feedback from an instrument. The outputs, at this level of granularity within their framework, are signals such as captured gestures that are used directly for performance feedback. At a higher level of abstraction is the sign or rule-based domain, which they equate to the ordering of pre-recorded or predetermined sections into sequences of events. Finally, at the highest level of abstraction within the framework is the symbol or knowledge domain, within which the musician is required to actively engage in interpreting abstract concepts (derived from scores or other abstract instructions). The power of this framework is that it describes a space within which it is possible to locate guidelines, principles, and models that have been derived from earlier DMI design and research, placing them within a context where the relationships between their different roles in the design process can be represented. At the signal level, for example, it is possible to locate the body of work on parameter mapping, whereas much work on characterizing expressive control would seem to belong to the symbolic or knowledge layer.

Taxonomies

Taxonomies are often used in HCI as a means of categorizing methods of design or evaluation according to characteristics that they have in common. The metaphor and embodiment taxonomy presented by Fishkin (2004), for example, was developed as a means of determining the "tangibility" of a tangible user interface (TUI), while providing a design space within which instances of TUIs can be compared in terms of their design metaphor, on one dimension, and their level of "embodiment," on the other dimension. With respect to DMIs, Paine (2010) has recently conducted a community-wide survey to create a comprehensive taxonomy of real-time interfaces for electronic music performance. Though at a preliminary stage, the process has already cast a spotlight on some interesting issues relating to how DMIs can be categorized, suggesting that categories such as "gestural controller," "digital controller," and "instrument" may represent a first level of division, but also pointing out that DMIs represent a new class of musical instruments not seen before in music history, in which the roles of creating real-time content and controlling the instrument are combined (Paine 2010). It should be noted that taxonomies differ subtly from frameworks in that their primary purpose is to define categories that are mutually exclusive, whereas frameworks tend to posit relationships that may be somewhat more interconnected.

In summary, frameworks often function to make explicit relationships between elements of an underlying theoretical approach, wrapping a design context around guidelines, principles, or even

models that instantiate this approach. Taxonomies function to categorize design approaches and, in doing so, can provide a basis for comparing elements from different categories along one or more salient design dimensions. Models are typically used to predict, for a given set of input conditions, how a system will perform. Principles represent general design goals, expressed in terms which provide little explicit direction as to how they should be implemented. Guidelines, on the other hand, are usually practical suggestions relating to how particular design goals might be achieved.

A Framework for Approaching the Evaluation of DMIs

Having presented some of the terminology used to contextualize the results of evaluations, let us return to the discussion of what kinds of questions different stakeholders might seek to answer through evaluation and why. As noted earlier, an HCI framework is usually a conceptual scaffold whose purpose is to set out the relationship between different approaches to the development of human–computer interfaces, whether that be in terms of theoretical foundation, design, or evaluation of design. The framework proposed subsequently addresses the evaluation of DMI designs from a number of perspectives that, in turn, reflect the roles of various stakeholders in the design process. Whereas the designer of a DMI may initiate the development of a new instrument, feedback from performers and audiences observing performances can play an important role in shaping the instrument as it evolves. The development of performance practice and a dedicated instrumental repertoire can go hand-in-hand with the evolution of a DMI, so that performers and composers become participants in shaping the function, form, and sound world of the instrument. Some instruments, such as Max Mathews's "Radio Baton" (1991), begin life as prototypes; yet, through repeated use in performance and the development of dedicated repertoire, these prototypes are refined and made robust to a point where they eventually became commercial products. Before setting out this framework, therefore, it is necessary to discuss evaluation from the perspective of each of these stakeholders, and to briefly summarize each perspective as it is currently represented in the DMI literature.

Evaluation of Performance—The Audience's Perspective

As many authors have pointed out, the greatest challenge facing designers of DMIs is that there is no longer a perceivable causal link between the gestures required to play the instrument and the mechanism that produces its sound. Schloss (2003), in particular, suggests that this disappearance of the relationship between cause and effect ultimately impacts upon the relationship between a performer and their audience. He notes that for approximately 30 thousand years, the way in which instruments produce sound has been physically evident to an audience. Only in the last 30 years or so has this relationship been dismantled. If, as Schloss implies, there is no reason for a causal relationship between gesture and sound to exist, how can the performer reconstruct a meaningful relationship between their actions and the sound that comes out of their instrument? Although he does not provide a solution to this problem, Schloss suggests that providing visual cues linking cause and effect, whether supported by the instrument or constructed as part of the performance (almost like choreography or magic), is a key component in making the performance convincing and effective. For this reason, he argues, people who perform should be performers and not merely "babysitters" of their sound-producing technologies.

Several studies indicate that observers of musical performances derive significant information about a performer's musical intent from observing how the performer moves during a performance. Davidson (1994), for example, asked several pianists to play the same piece in three ways—once with as little expression as possible, once projecting their expressive intent, and once with exaggerated expression. Performances were recorded using Johansson's point-light technique (Johansson 1973) and then replayed to a group of naive observers under three

conditions—sound only, point-light video only, or sound and video together. They were then asked to rate performances according to a seven-point scale from deadpan to exaggerated. Results clearly demonstrated that participants were able to consistently recognize a performer's expressive intent, regardless of whether they could see, hear, or both see and hear the performance. It is significant that the observers' mean average ratings ranged most widely from deadpan to exaggerated conditions when both audio and visual information were present. Later studies have confirmed this finding. Dahl and Friberg (2007), in a study exploring the communication of different emotional intentions in marimba, bassoon, and saxophone performances, found that participants were able to distinguish between happiness, sadness, and anger at levels above chance when presented with only video footage. It is interesting to note that all of these studies, and indeed many others supporting Davidson's findings, have used performances on acoustic instruments. Perhaps a valid evaluation of DMI designs might be to see if, given the same task, audiences were similarly consistent in their responses.

In contrast to the work of authors such as Davidson and Friberg who focus on how well performers communicate expressive nuance to their audience, Fyans, Gurevich, and Stapleton (2009) have begun to develop a model of a spectator's understanding of error in performance based in part on their understanding of performer intent. Schloss (2003) has noted that people attend musical performances in part to observe skilled players doing something they cannot do themselves. But what is it about the interaction between a performer and their instrument that enables audience members to make this judgment? Fyans, Gurevich, and Stapleton suggest that, for the spectator, skill is a judgment based on a combination of knowledge, experience, and their own assessment of the degree of difficulty of a piece. They suggest that the spectator's understanding of success is a continuous measure of the distance between the performer's intent (as understood by the spectator) and the performance outcome (as perceived by the spectator). The problem posed by DMIs is that there is often no way for a spectator to have acquired knowledge about how an instrument should be played: DMIs often have no associated repertoire or performance history and often no obvious metaphor connecting them to prior musical contexts. The challenge, therefore, is to determine how audiences disentangle judgments about performance error from judgments concerning instrument failure—both result in a breakdown between a performer's intent and the outcome of an action, but the source of the former is a mistake by the player, whereas the source of the latter is a failure of the technology. Fyans, Gurevich, and Stapleton, adapting the framework for designing interactive sensing systems proposed by Bellotti et al. (2002), suggest that designers of DMIs can, as part of their design process, obtain feedback from performance spectators by posing similar questions, e.g: (1) How does the spectator identify the intended target for an action? And (2) How does the spectator know what to ignore in observing a performance?

In summary, evaluating DMIs from an audience's perspective presents challenges that are unmatched in traditional performance contexts. It is therefore up to the performer and instrument designer to help the spectator by reintroducing causal relationships that allow for the modeling of performer intent. Moreover, involving the spectator in the process of designing a DMI can ensure that such intended causal links are evaluated by an important stakeholder (a potential audience member) at an early stage in the design process.

Evaluation through Practice—The Performer's Perspective

There is no doubt that the most important stakeholder in the process of designing and building a DMI is the performer. Unless the instrument can successfully translate their musical intent into sound in a reliable way it fundamentally fails as an instrument. Even in performances where the intent of the composer and/or performer is for the outcome of a process to be unpredictable, such as in cases where chaotic or stochastic elements are introduced as part of the synthesis or compositional process, these too must behave in a reliably unpredictable way. It is not surprising, therefore, that

much evaluation of DMI design has been rightly focused on determining how well instruments meet the needs of performers, and of audiences as the observers of performances. Although ultimately the best evaluation of a performance is one's own impression of how compelling it was to participate in or to attend, experienced computer music practitioners can provide us with guidelines or principles. These, in turn, are invaluable in informing the design of new DMIs.

The first point, which emerges from a number of such reflections, is that DMI players are not stupid; their instruments should be challenging to master and thereby engage them in developing virtuosity. David Wessel and Matthew Wright (2001), reflecting on many years of performance experience with DMIs, suggest that from the performer's perspective, such musical instruments succeed or fail for a number of reasons. Firstly, there are sociological reasons such as lack of a developed repertoire. Secondly, there are practical factors such as the reactive behavior of the instrument (e.g., the presence of latency and jitter that perturb fluent interaction with the instrument), the lack of coherence of the cognitive model underlying the design, the ease of use, and the potential for development of virtuosity. Dobrian and Koppelman (2006) argue that designers of DMIs can also benefit from studying acoustic instruments. Such studies can help develop relationships between playing gestures and sound that are both complex and intuitive, that intelligently characterize performance gestures, and that take advantage of existing instrumental skills. However, as Cook (2001, p. 1) argues, "copying an instrument is dumb," but finding a way to leverage expert technique makes sense, as you can thereby take advantage of all the years of practice that the performer has already invested in their technique. In building upon an existing instrument metaphor, though, Cook also points out that not all players have the ability to incorporate additional functionality, so that simply adding sensors to an instrument and expecting a performer to incorporate the gestures required to integrate them into their playing will not always work: "some players have spare bandwidth, some do not." Further, as the pianist Sarah Nicolls points out, adding sensors to an instrument not only changes its functionality, but can also fundamentally alter the playing techniques required to achieve mastery. In reflecting upon performing with her augmented piano, Nicolls writes,

> "Imagine the pianist lifting the arm away from the keyboard, perhaps signifying a breath between musical phrases. By using this gesture to generate data and in turn the processing of sound, I found, in making such a gesture, I was now focused on playing the sensors and NOT the previously almost subconscious movement—thereby turning the gesture into a material action. As a solo performer is only one body, one mind, these cycles of complexity and confusion in fact perhaps begin to disrupt the artistic spontaneity and intuitive physical sense and the original meaning of the gesture is potentially undermined" (Nicolls, 2010, p. 50).

Nicolls thereby cautions both the instrument builder and those composing for augmented instruments to consider such basic aspects of playing as the natural ebb and flow of effort in performance and the player's need to "recover" within the flow of a piece as a crucial part of sustaining skilled performance.

In summary, performance should be considered as the ultimate evaluation of any instrument design, and digital instruments are no exception. Performers are the only people who can provide feedback on an instrument's functioning in the context for which it was ultimately intended, that of live music making. And yet performers, too, can adapt to properties of instruments that are non-ideal—the sticky pedal on a piano, for example—so that an impartial assessment of an instrument's playability is also desirable if a solid design is to be assured.

Evaluation of Interaction—The Designer's Perspective

From the foregoing discussion, it becomes apparent that more traditional evaluation methodologies for human–computer interaction are, in many cases, unsuited to the evaluation of DMIs. In fact, as Bellotti et al. (2002) point out, such methodologies

are inappropriate for a whole class of interactive sensing systems. Traditional models for interaction design in HCI, such as Norman's action-execution model, place the responsibility for figuring out what state the system is in, and what actions are currently possible, firmly in the hands of the user (Norman 1988). Bellotti et al. (2002) suggest that, in comparison, interactive sensing systems more closely resemble the paradigm of person-to-person communication: Both parties must be able to establish a shared topic, such as a shared action goal. Significantly, in interactive sensing systems, both the system and the user operate within the same timeframe so that actions performed by one can be immediately coupled to responses of the other.

However, such a tight coupling between action and response presents difficulties in determining how to evaluate a user's experience, because interrupting their actions inevitably interrupts their thoughts and the achievement of their musical goals. Any comment on the behavior of the system is, therefore, made off-line with respect to the task being evaluated.

Evaluating Playability

It is often necessary to probe interaction designs at the task level, particularly in order to evaluate two possible options for a given design, or to probe the mental model that a user is constructing of a given interaction task. In terms of a methodology, Wanderley and Orio (2002) have provided a contribution for evaluating the usability of DMIs. They recommend that evaluations should be constructed around simple musical tasks, such as reproducing musical units like glissandi or arpeggios. Furthermore, they suggest that evaluations should be placed within musical contexts or metaphors such as note-level control, score-level control, or sound-processing control (post-production activity) in order to aid the user in constructing an appropriate model of the task. In choosing appropriate tasks for such usability evaluation, they suggest that relevant features to be tested might include learnability, explorability, feature controllability, and timing controllability.

Several authors have subsequently implemented this framework in constructing evaluations of instruments and controllers. Poepel (2005), for example, constructs a set of simple musical tasks and associated indicators according to Wanderley and Orio's framework. These tasks are constructed from skills known to lie within both the skill of the string players who were his participants, and the capabilities of the controller he was evaluating. The indicators used to evaluate the system, with respect to its ability to support expressive playing, were operationalized as tasks which were then grouped into categories. Timing accuracy, for example, incorporated a group of tasks characterized by tempo, timing, and pauses, and was represented in Poepel's study by tasks such as pizzicato, collé, spiccato, and playing short notes. What is particularly interesting here is that Poepel not only applies Wanderley's guidelines to the design of the evaluation, but also finds a way to map elements of advanced string technique onto musical tasks that, while modular in nature, still allow the player to probe the expressive potential of the interface.

An alternative framework for evaluating DMIs for their potential to support performers is proposed by Jorda (2004). In endeavoring to capture Wessel's notion of the need for DMIs to support the acquisition of skill and thereby the engagement of skilled performers over a long period of time (Wessel and Wright 2001), Jorda suggests that DMIs, indeed all instruments, can be described in terms of their ability to support diversity in musical style and performance. He therefore addresses similar issues in terms of evaluating an instrument's ability to support a performer in realizing musical goals, but frames his discussion not in terms of individual tasks, but classes of tasks that are expressed as levels of *diversity*. Jorda defines three classes in his framework, Macro, Mid, and Micro, and suggests that instruments can be described in terms of the "level" to which they support these diversity goals. Macro diversity ("MacD"), or stylistic diversity, refers to an instrument's ability to be used in a wide variety of musical contexts or styles, with more "general-purpose" instruments such as the harmonica having a higher level of MacD than those such as the double bass. Instruments with a high level of MacD, he suggests, are relatively easy to play for an amateur and are well suited

to autodidactic methods. Mid diversity ("MidD"), or "performance diversity," captures the degree to which two performances on the same instrument can differ. Instruments capable of supporting a large repertoire of music requiring a high level of skilled performance, such as a violin, have a high level of MidD, whereas sound installations, which always produce the same sound in response to a limited repertoire of actions, have a low level of MidD. Micro diversity ("MicD"), captures the degree to which an instrument can reflect very fine nuances in performance. Instruments with a high level of MicD are those which are highly responsive to tiny nuances in performance gestures, such as a violin. Instruments with a low level of MicD do the same thing regardless of the way in which they are played—e.g., pressing a single button to start a sequencer. Instruments with high levels of MicD, Jorda suggests, have the greatest potential to encourage the acquisition of virtuosity and the development of sophisticated musical repertoire. It is important to understand that Jorda's diversity classes are not mutually exclusive—instruments that support performance diversity (high MidD) are also likely to support subtlety in performance nuance (i.e., they will also have a high level of MicD).

Although other approaches to evaluating DMIs at the task level exist, that of Wanderley and Orio has gained a firm foothold within the DMI design community. It is important to remember, however, that this framework is intended to evaluate the usability of DMIs, and that there exist many other methods, such as that of Jorda, that provide feedback on other aspects of DMI designs.

Evaluating Playing Experience

The challenge in carrying out an evaluation of performance on any musical instrument is in finding a way to probe the player's experience without disrupting their engagement with the task. For traditional HCI evaluations, such as those of desktop interfaces, a number of techniques have been developed where a user's experience is captured as they work. Some of the most successful of these involve so-called "think-aloud" protocols where users provide a running commentary on their thought process as they navigate the interface (van den Haak and de Jong 2003). But such techniques are of little use during any substantive musical task: They require the player to speak while they are both performing and evaluating the instrument, a task which is virtually impossible. Strain and colleagues have recently presented a modified think-aloud protocol that overcomes some of these limitations, where participants are asked to reflect on their experience after the event, but are prompted by the experimenter, who reminds them of actions they performed or problems they appeared to encounter during their interaction with the system (Strain, Shaikh, and Boardman 2007). Though this method was developed for the evaluation of speech output systems for blind computer users, it has recently been adapted for use in the evaluation of a percussion controller (Chuchacz 2009). These scenarios have much in common, as in both cases normal think-aloud protocols are inappropriate because the user is fully engaged in attending to the audio output of the system.

Because what is ultimately most important is how suitable a DMI is for music making in a live performance context, any evaluation must be able to assess how well the instrument performs in terms of both its practical and musical usability. Most evaluations reported in the DMI literature achieve this goal by using qualitative and quantitative methods in parallel. Such studies typically capture real-time sensor data from an instrument that are then analyzed with respect to measures such as accuracy in timing or the trajectory of movements in order to judge some aspect of the instrument's controllability. Qualitative data in the form of responses to questionnaires are often gathered to probe the player's experience of the instrument at a cognitive level, and both data sets are then used in evaluating the system. The result is that the player's performance on the system, and their experience of performing on the system, are not assessed at the same time—the questionnaire typically records the player's reflections upon their experience after the event. In an extension of this paradigm, some authors have employed more advanced lexical analysis tools, such as discourse

analysis and concept mapping, to quantify data obtained through semi-structured interviews (Paine, Stevenson, and Pearce 2007; Stowell, Plumbley, and Bryan-Kinns 2008; Chuchacz 2009). When compared to choice-based questionnaires, this approach has the advantage of being able to capture subtle nuances in participants' responses to probe questions and to allow quantitative results to be extracted from unstructured dialogue.

Is it possible to devise a methodology that could capture a player's experience of playing while they are engaged in a musical task? The growing trend in HCI research toward evaluating experience rather than efficiency is currently giving rise to new methods for evaluation that are experience- rather than task-focused (Kaye 2007). This need has arisen because HCI now includes a whole class of interactions that, like musical performance, couple real-time motion capture with real-time audio, video, and often haptic feedback. The evaluation of environments for immersive gaming and training, for example, requires the development of protocols that can capture a user's experience while interacting within a fully immersive, time-critical context. As Kiefer, Collins, and Fitzpatrick (2008) suggest, ultimately the best way to do this may be by using non-intrusive systems to monitor physiological data such as electroencephalography, electromyography, and galvanic skin response sensing. While physiological tracking has already been employed as an input method for musical control (Knapp and Cook 2006; Coghlan and Knapp 2008), it does not yet appear to have been used directly to evaluate the experience of playing a DMI.

HCI research on the evaluation of interactive game playability is relevant in a discussion of the experience playing DMIs. Although at their simplest these techniques incorporate fairly standard usability measures, such as time-on-task, etc., the evaluation of so-called "long games" has generated methods that appear to have potential for evaluating the learnability and playability of new musical instruments. "Long games" are those defined by the industry as having single playing sessions that last for approximately one hour but that have a maximum playing length of tens of hours. As such, engagement with long games has much in common with learning to play a new musical instrument or learning to play a new piece on an instrument that has already been mastered. Febretti and Garzato (2009) have measured the correlation between usability and playability factors for long-term user engagement in eight commercial "long games." Using a combination of heuristic techniques and data captured through game rating aggregators available on specialized Web sites for game quality assessment, they determined that long term engagement is more significantly affected by the density of usability defects for a given gaming session than by the overall number of such defects in the game. If the density of usability defects is low, i.e., a particular usability problem crops up infrequently during a gaming session, the effect on engagement is temporary and local. If, on the other hand, the density of usability defects is high, playing experience and engagement are significantly affected. Though such a study does not appear to have been carried out in the context of digital instrument evaluation, these findings suggest that the number of usability defects encountered within a given performance is more likely to have an impact on playing experience than the total number of usability problems encountered with any one instrument across its lifetime. This is certainly a hypothesis that deserves to be tested.

Evaluation within the Marketplace—The Manufacturer's Perspective

Although a discussion of marketing of DMIs falls outside the scope of the current article, it is important to remember that manufacturers have a crucial role to play in the development of DMI designs. The components upon which designers rely are themselves developed and tested by companies who are thereby invested in the success of a particular DMI as a product. Other companies exist who create middleware such as plug-and-play hardware, synthesis software, etc., and these companies must be considered as stakeholders in the design process because they too are interested in the success of the elements that they supply. Indeed, the development of tools and components can greatly shape

Table 1. Methods Used by Different Stakeholders for Evaluating DMI Designs

	Possible Evaluation Goals			
Stakeholder	*Enjoyment*	*Playability*	*Robustness*	*Achievement of Design Specifications*
Audience	critique, reflection, questionnaires, observational studies	experiments concerning mental models		
Performer/ Composer	reflective practice, development of repertoire, long-term engagement (longitudinal study?)	quantitative methods for evaluation of user interface, mapping, etc.	quantitative methods for hardware/ software testing	
Designer	observation, questionnaire, Informal feedback	quantitative methods for user interface evaluation		use cases, feedback regarding stakeholder satisfaction
Manufacturer	market surveys, sales	sales, consumer feedback	quantitative methods for hardware/ software testing, consumer feedback	market penetration (performers, consumers), sales, consumer feedback

the direction of DMI design (for example, consider the adoption of general purpose interfaces such as the WIImote, the Wacom tablet, and environments such as Supercollider and MaxMSP). There is also an iterative element—products which are infrastructure for today's DMIs were themselves once prototype systems in their own right, and are robust and successful because they too were the result of an iterative process of design and evaluation. Therefore, manufacturers must be considered as stakeholders who are engaged with DMI design and evaluation, not just at the end of a cycle of development, but also within the lifetime of most, if not all, DMI designs.

A Proposed Evaluation Framework

From this discussion, it is apparent that the design of a DMI is dependent upon input from many stakeholders, each of whom brings to the table their own means of evaluating the instrument from their perspective. Initial design decisions concerning aspects such as interaction or sound synthesis are associated with very specific methods for assessing success, e.g., benchmarking against existing models or standards for the relevant discipline. At some stage, however, the instrument ceases to be a collection of systems and becomes an entity in its own right, an entity whose function is to translate the intent of a performer into music, music which in turn will be experienced by an audience. Thus it is useful to develop a means of relating the interests of different stakeholders on one hand to the variety of reasons for evaluating instruments on the other, in order that the relationship between the two can be more readily understood.

The framework proposed in Table 1 is an attempt to accomplish this task. It is, as most frameworks are, primarily a scaffold, designed to provide a space within which these relationships can be further explored and within which a given evaluation of a DMI can be situated. It is populated here not with particular DMI evaluations, but with methods that represent the kinds of evaluations that have been carried out by a given stakeholder group to evaluate

an instrument with respect to a given goal. Table 1 illustrates possible stakeholders in the process of designing and evaluating DMI designs on the left, with possible goals for such evaluations on the right. The entries in the table represent methods that a given stakeholder might use in order to evaluate a DMI against a given design goal. Any single DMI may be the subject of more than one evaluation in the process of its design cycle, and these can easily be situated and related to each other within this grid.

Although the foregoing discussion has been organized with respect to stakeholder interests as represented by the left-hand column in this table, it is equally possible to approach a discussion of evaluating DMIs from the perspective of design goals such as those indicated on the right. Indeed, the discussion of playability already addresses the evaluation of DMIs with respect to the first of these goals. The literature documenting evaluation of interaction in HCI is rarely framed in terms of goals such as "enjoyment," "robustness," or "achievement of design specifications." In the case of goals such as robustness, which largely become issues at the product development phase, evaluation is typically closely tied to component design, which in turn is bound up with the intellectual property of an individual or company and as such is information that is often not available within the public domain. Similar issues may arise concerning the evaluation of instruments against design specifications when these specifications are commercially sensitive. However, it is entirely possible for an instrument to emerge as a result of a happy accident—consider the origin of the Theremin, which was a by-product of the development of early radio transmission technology (Glinsky 2000). In practice, most DMIs are developed in response to some design decision, however loosely stated—a decision that serves to provide some practical design constraints as the instrument evolves (see, for example, Blaine and Fels 2003).

The concept of enjoyment has, however, received some considerable interest within HCI, particularly as it relates to the evaluation of gaming experience. Sweetser and Wyeth (2005) have developed a model for evaluating enjoyment of game play in terms of "game flow." The model they propose consists of eight elements—concentration, challenge, skills, control, clear goals, feedback, immersion, and social interaction. When used to evaluate two games which had previously been rated as having high and low levels of enjoyment using a preexisting tool, their model was successful in drawing out those aspects of the games that contributed to how enjoyable they were to play. As many of these elements find their parallel in musical performance, a model such as this might be adapted to evaluate enjoyment with respect to learning and playing DMIs.

In reflecting upon this framework, certain aspects of the process of evaluating DMI designs begin to emerge. For example, though both performer's and audiences are interested in modeling the causal link between the actions performed upon an instrument and the sound that results, the methods by which they might assess an instrument along this dimension may be quite different. Performers require stability and reliability from their instrument in order to model this causality, and this might be measured at the level of specificity, such as system jitter and reliability of sensor response (Wessel and Wright 2001). Audiences, as discussed earlier, also need to model causality, but at a much more cognitive level, and their understanding is typically explored using tools to probe such cognitive models. Plugging a specific DMI into this framework, one can imagine that considerations of causality for a performer will most likely be addressed at a stage of the design that precedes any consideration of audience evaluation.

Another interesting aspect of the framework proposed herein is the content of the cells in Table 1. These represent the points of intersection between stakeholder interests and design goals with respect to evaluation. Although some of these points of intersection, such as between Performer/Composer and Playability, have been extensively explored through studies such as those cited herein, others, such as the intersection between Audience and Enjoyment, are empty. In other words, the table highlights a somewhat uneven distribution of evaluation studies, with a greater emphasis on types of evaluation that are well supported by methodologies from HCI in general and from

the DMI design literature in particular. What this preliminary sketch of the framework reveals, therefore, are the opportunities and challenges that remain, both in terms of developing appropriate methodologies for evaluation for these areas, and for carrying out such evaluations. One important caveat is that not all of the points of intersection here may need to be addressed—some stakeholders simply may not have any interest in evaluating a DMI against a particular design goal. Audiences, for example, may not be concerned directly with an assessment of how closely an instrument design maps onto its initial design specification. Other empty squares on this grid clearly represent open questions—does a manufacturer currently have any processes in place for evaluating the enjoyability of their instrument? In short, there are likely to be many more possible goals for evaluation than those stated here, and new stakeholders in the design process are likely to emerge in time. This scaffold functions merely as a starting point for further discussion of the role of evaluation in the design of DMIs.

Conclusion

As stated at the outset of this article, the purpose of this discussion has been to reflect upon the role that evaluation plays in the process of designing DMIs. The starting point for this discussion was an acknowledgement that there are many stakeholders in this design process, each of whom may have different goals with respect to evaluating an instrument, and different methodologies that they bring to the evaluation process. In fact, their different perspectives are all necessary, but often at different phases of an instrument's design cycle. It is important, however, for all involved in the process of designing DMIs to have a shared understanding of the role of any given evaluation, and to have a clear understanding of the questions being posed and the stakeholder perspective from which they have emerged. Only then can the outcome of such an evaluation be placed in perspective within the overall design process. The DMI evaluation framework presented herein is an attempt to provide some conceptual scaffolding within which both the interests of stakeholders in the design process and possible goals of evaluation might coexist. The goal is to provide a means for those involved in DMI design to understand how these different perspectives contribute to the creation of the final product—an instrument that engages both performer and audience alike.

References

Bellotti, V., et al. 2002. "Making Sense of Sensing Systems: Five Questions for Designers and Researchers." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our world, changing ourselves.* New York: Association for Computing Machinery, pp. 415–422.

Blaine, T., and S. Fels. 2003. "Contexts of Collaborative Musical Experiences." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME).* Montreal, Canada: McGill University, pp. 129–134.

Choi, I., R. Bargar, and C. Goudeseune. 1995. "A Manifold Interface for a High Dimensional Control Space." In *Proceedings of the International Computer Music Conference.* San Francisco, California: International Computer Music Association, pp. 181–184.

Chuchacz, K. 2009. "Real-time, Hardware Implementation and Musical Interface Design for a Percussion Instrument Based on a Physical Model." PhD Thesis, Queens University Belfast.

Coghlan, N., and R. B. Knapp. 2008. "Sensory Chair: A System for Biosignal Research and Performance." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME).* Genova: Casa Paganini, pp. 233–236.

Cook, P. 2001. "Principles for Designing Computer Music Controllers." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. New York: Association for Computing Machinery, pp. 1–4.

Cook, P. 2009. "Re-Designing Principles for Computer Music Controllers: A Case Study of SqueezeVox Maggie." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Pittsburgh, Pennsylvania: Carnegie Mellon University, pp. 218–221.

Cooper, A. 1995. *About Face—The Essentials of User Interface Design*. Foster City, California: IDG Books Worldwide.

Dahl, S., and A. Friberg. 2007. "Visual Perception of Expressiveness in Musicians' Body Movements." *Music Perception* 24(5):433–454.
Davidson, J. W. 1994. "What Type of Information is Conveyed in the Body Movements of Solo Musician Performers?" *Journal of Human Movement Studies* 6:279–301.
Dobrian, C., and D. Koppelman. 2006. "The 'E' in NIME: Musical Expression with New Computer Interfaces." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Paris: Institut de Recherche et Coordination Acoustique/Musique, pp. 277–282.
Dumas, J. S., and J. C. Reddish. 1999. *Practical Guide to Usability Testing, 2nd ed*. Chicago, Illinois: Chicago University Press.
Febretti, A., and F. Garzotto. 2009. "Usability, Playability, and Long-Term Engagement in Computer Games." In *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems*. New York: Association for Computing Machinery, pp. 4063–4068.
Fishkin, K. 2004. "A Taxonomy for and Analysis of Tangible Interfaces." *Personal and Ubiquitous Computing* 8(5):347–358.
Fyans, A. C., M. Gurevich, and P. Stapleton. 2009. "Where Did It All Go Wrong? A Model of Error From the Spectator's Perspective." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Pittsburgh, Pennsylvania: Carnegie Mellon University, pp. 172–175.
Gelineck, S., and S. Serafin. 2009. "A Quantitative Evaluation of the Differences between Knobs and Sliders." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Pittsburgh, Pennsylvania: Carnegie Mellon University, pp. 13–18.
Gentner, D., and A. L. Stevens, eds. 1983. *Mental Models*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Glinsky, A. 2000. *Theremin: Ether Music and Espionage*. Champaign, Illinois: University of Illinois Press.
Hornecker, E., and J. Buur. 2006. "Getting a Grip on Tangible Interaction: A Framework on Physical Space and Social Interaction." In *Proceedings of the 2006 Computer-Human Interaction Conference*. New York: Association for Computing Machinery, pp. 437–446.
Hunt, A., and R. Kirk. 2000. Mapping Strategies for Musical Performance. In M. Wanderley and M. Battier, eds. *Trends in Gestural Control of Music*. Paris: Institut de Recherche et Coordination Acoustique/Musique, pp. 231–258.
Johansson, G. 1973. "Visual Perception of Biological Motion and a Model for its Analysis." *Perception and Psychophysics* 14(2):201–211.
Jorda, S. 2004. "Digital Instruments and Players: Part II: Diversity, Freedom and Control." In *Proceedings of the International Computer Music Conference.* San Francisco, California: International Computer Music Association, pp. 706–710.
Kaye, J. 2007. "Evaluating Experience-Focused HCI." In *Proceedings of the 25th International Conference Extended Abstracts on Human Factors in Computing Systems*. New York: Association for Computing Machinery, pp. 1661–1664.
Kiefer, C., N. Collins, and G. Fitzpatrick. 2008. "HCI Methodology For Evaluating Musical Controllers: A Case Study." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME).* Genova: Casa Paganini, pp. 87–90.
Knapp, R. B. and P. R. Cook. 2006. "Creating a Network of Integral Music Controllers." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Paris: Institut de Recherche et Coordination Acoustique/Musique, pp. 124–128.
Malloch, J., D. Birnbaum., E. Sinyor., and M. M. Wanderley. 2006. "Towards a New Conceptual Framework for Digital Musical Instruments." In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06).* Montreal: McGill University, pp. 49–52.
MacKenzie, I. S. 2003. "Motor Behaviour Models for Human–Computer Interaction." In J. M. Carroll, ed. *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*. San Francisco, California: Morgan Kaufmann, pp. 27–54.
Mathews, M. 1991. "The Radio Baton and Conductor Program, or: Pitch, the Most Important and Least Expressive Part of Music." *Computer Music Journal* 15(4):37–46.
Nicolls, S. 2010. "Seeking Out the Spaces Between: Using Improvisation in Collaborative Composition and with Interactive Technology." *Leonardo Music Journal* 20:47–55.
Norman, D. 1988. *The Design of Everyday Things*. New York: Doubleday/Currency.
Paine, G. 2010. "Towards a Taxonomy of Realtime Interfaces for Electronic Music Performance." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Sydney: University of Technology Sydney, pp. 436–439.
Paine, G., I. Stevenson, and A. Pearce. 2007. "The Thummer Mapping Project (ThuMP)." In *Proceedings of the International Conference on New Interfaces*

for Musical Expression (NIME). New York: New York University, pp. 70–77.
Poepel, C. 2005. "On Interface Expressivity: a Player-Based Study." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Vancouver: University of British Columbia, pp. 228–231.
Rasmussen, J. 1986. *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. Amsterdam: Elsevier Science Inc.
Schloss, W. A. 2003. "Using Contemporary Technology in Live Performance: The Dilemma of the Performer." *Journal of New Music Research* 32(3):239–242.
Serafin, S. 2004. Physical Synthesis of Bowed String Instruments. In K. Greenebaum and R. Barzel, eds. *Audio Anecdotes III: Tools, Tips, and Techniques for Digital Audio*. Natick, Massachusetts: AK Peters, pp. 85–98.
Smith, J. O. 2008. "Physical Audio Signal Processing, December 2008 Edition." Available on-line at ccrma.stanford.edu/~jos/pasp/. Accessed 11 October 2009.
Stowell, D., M. Plumbley, and N. Bryan-Kinns 2008. "Discourse Analysis Evaluation Method for Expressive Musical Interfaces." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. Genova: Casa Paganini, pp. 81–86.
Strain, P., A. D. Shaikh, and R. Boardman. 2007. "Thinking But Not Seeing: Think-Aloud for Non-Sighted Users." *Proceedings of the 25th International Conference Extended Abstracts on Human Factors in Computing Systems*. New York: Association for Computing Machinery, pp. 1851–1856.
Sweetser, P., and P. Wyeth. 2005. "GameFlow: a Model for Evaluating Player Enjoyment in Games." *Computers in Entertainment* 3(3):1–24.
van den Haak, M., and M. de Jong. 2003. "Exploring Two Methods of Usability Testing: Concurrent Versus Retrospective Think-aloud Protocols." In *Proceedings of 2003 IEEE International Professional Communication Conference*. Orlando, Florida: IEEE Professional Communication Society, pp. 285–287.
Verplank, W. 2003. *Interaction Design Sketch Book*. Unpublished paper. Available on-line at www.billverplank.com/IxDSketchBook.pdf. Accessed 16 July 2010.
Wanderley, M. M., and N. Orio. 2002. "Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI." *Computer Music Journal* 26(3):62–76.
Wessel, D., and M. Wright. 2001. "Problems and Prospects for Intimate Musical Control of Computers." In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. New York: Association for Computing Machinery, pp. 1–4.