Methodological Approaches to the Evaluation of Game Music Systems

Anthony Prechtl
Department of Computing
The Open University
Milton Keynes, UK
anthony.prechtl@open.ac.uk

Robin Laney
Department of Computing
The Open University
Milton Keynes, UK
r.c.laney@open.ac.uk

Alistair Willis
Department of Computing
The Open University
Milton Keynes, UK
a.g.willis@open.ac.uk

Robert Samuels
Department of Music
The Open University
Milton Keynes, UK
robert.samuels@open.ac.uk

ABSTRACT

Despite an emerging interest in the application of dynamic computer music systems to computer games, currently there are no commonly accepted approaches to empirically evaluating game music systems. In this paper we pose four questions that researchers could assess in order to evaluate different aspects of a game music system. They focus on the music's effect on the game playing experience (whether the music leads to a more enjoyable experience, and whether it affects the player in the intended way during the game), and how the music itself is perceived (whether it reaches a certain aesthetic standard, and whether it accurately conveys the intended narrative). We examine each of these questions in turn, for each one establishing a theoretical background as well as reviewing and comparing relevant research methodologies in order to show how it could be addressed in practice.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Methodologies and techniques*; K.8.0 [**Personal Computing**]: General—*Games*

General Terms

Experimentation, Theory, Verification

1. INTRODUCTION

Recently there has been a growing interest in the use of dynamic computer music systems to produce background music for computer games. For example, researchers have proposed and developed systems that, in response to game events, algorithmically generate music in real time [7, 36, 13, 5, 34], or intelligently crossfade different prerecorded pieces of music [25]. However, only a few (e.g., [36]) have included an empirical evaluation component. As Nierhaus [29] points out, there is a tendency among algorithmic music researchers to focus purely on the architecture of their systems, with the actual musical output serving mainly as a confirmation of system functionality. Pearce et al [32] further argue that the lack of empirical evaluation in the development of algorithmic music systems has "compromised the practical or theoretical value" (p. 120) of the research. This tendency is largely paralleled in game music research, and has arguably led to a situation in which there are several interesting and potentially relatable dynamic game music systems, but no obvious ways to compare them, or even to consider them under a unifying set of principles.

The absence of a standard methodology for evaluating game music systems could be due to multiple factors. First, there are many different genres of games, and even significant variation within genres, as well as different styles of game playing. This means that overall there is relatively little generalizability across game experiences. Additionally, game playing involves complex multi-modal interaction which, despite receiving a surge of academic interest over the past decade or so, is still not particularly well understood. Indeed, barring commercial performance after a game's release, it is not obvious what specifically it means for a game to be successful, let alone how to empirically measure success.

Nonetheless, game researchers have proposed several useful methodological approaches for empirically evaluating different aspects of games and game playing experiences. These approaches loosely fall into one of two categories: Some aim to directly or indirectly examine player enjoyment, typically through the use of subjective questionnaires or structured interviews. Others focus on measuring players' psychophysiology as an indication of their emotional response during game play, under the assumption that what is being evaluated (in our case, the music) should affect it in a certain intended way. Both of these approaches are distinctly *player oriented* in the sense that they focus on the effect of the music on the player's experience, rather than how the mu-

AM '14, October 01 - 03 2014, Aalborg, Denmark

ACM 978-1-4503-3032-9/14/10 ...$15.00.
http://dx.doi.org/10.1145/2636879.2636906.

sic itself is perceived. In general, however, the first is more subjective and aesthetically motivated, whereas the latter is more objective and functionally motivated.

A similar distinction can be drawn in evaluation methodologies for music systems outside the domain of games. A key difference between these *music-oriented* approaches and player-oriented ones is that music-oriented approaches involve participants critically rating different aspects of the music itself, rather than the experience of playing the game and how the music contributed to it. From the subjective and more aesthetically-motivated perspective, one can evaluate whether the music reaches a certain aesthetic standard or, in a similar vein, is stylistically plausible. From the more objective and functionally-motivated perspective, one can evaluate whether the music accurately conveys the intended narrative.

In light of our distinction between player- and music-oriented approaches, and further, between aesthetic and functional ones, we pose four questions for the empirical evaluation of game music systems:

1. Does the music lead to a more enjoyable game playing experience?
2. Does the music affect the player in the intended way during game play?
3. Does the music sound aesthetically reasonable or stylistically plausible?
4. Does the music convey the intended narrative?

In this paper, we will discuss each of these questions in turn (Sections 2.1, 2.2, 3.1, and 3.2, respectively), for each one providing a theoretical foundation and reviewing related research. We will also describe what success might mean in each case, as well as possible limitations.

It is worth noting that other evaluation questions may also be relevant. These include technical considerations such as a system's processing and memory requirements, usability considerations such as how easy it is for a game designer to interface with, and possibly others. However, we consider these beyond the scope of this paper, in which we are concerned more with game music itself and its effects on player experience.

2. PLAYER-ORIENTED APPROACHES

Player-oriented approaches to game evaluation aim to capture aspects of a player's experience with a game. In the domain of game music, these approaches could have relatively high ecological validity since they inherently involve actual game playing rather than just listening to the music. In practice, however, it would be impossible to separate aspects of a music system itself from the way in which the game was configured to control it, which could easily lead to confounding.

For example, in the classic 1985 Nintendo game *Super Mario Bros.*, when the player begins to run out of time to complete the current level, the tempo of the background music increases by a certain amount, emphasizing the sense of urgency. Presumably, the game's internal logic is responsible for specifying to the music system when and by how much to modulate its tempo, while the music system is responsible for handling the command in a musically satisfactory way. Consider, then, a study participant rating the music negatively during the tempo change—this could equally be the result of poor handling of the tempo change command (for example, changing the tempo in the middle of a phrase), or the choice of an unfitting tempo. Thus, the underlying music system and the way the game was configured to control it would confound each other. Because of cases like this, when using a player-oriented approach it is important to be specific about which side of the system is being tested, and take steps to control the other as best as possible.

2.1 Subjective measures of player enjoyment

The first evaluation question we pose is "Does the music lead to a more enjoyable game playing experience?" Player enjoyment is arguably the single most important consideration in game design, so if the music could somehow increase it, that would certainly be a good indication of success. Simply asking participants how much they enjoyed a particular game condition, or which one they preferred, is certainly not out of the question, and has been done in [17] and [43], for example. However, it is not obvious how to codify enjoyment of a game playing experience in such a way that is clear to participants yet also conducive to rigorous analysis. Additionally, differences between experimental conditions may be subtle enough that a player could not consciously distinguish between them, or may have difficulty discerning a preference. Thus, many researchers have chosen to evaluate enjoyment indirectly—that is, to evaluate other aspects of the game playing experience which are more easily or clearly measured, under the assumption that they are indicative of enjoyment.

Perhaps the most common game enjoyment metrics are related to the concept of *flow*, first described by psychologist Mihalyi Csikszentmihalyi [8]. According to Csikszentmihalyi, flow is an optimal psychological state of strong enjoyment characterized by deep concentration and a loss of awareness of oneself and time. He notes that flow tends to occur during the performance of a task when a set of specific conditions are met, including, for example, that the task's difficulty matches the person's skill level, and that the task has clear goals and provides immediate feedback. Flow has been empirically evaluated and measured in a variety of ways—an overview is provided in [24].

Sweetser and Wyeth [39] proposed one of the earliest formal strategies for evaluating game player enjoyment, named *GameFlow* and based entirely on flow. They brought Csikszentmihalyi's conditions for the occurrence of flow into the domain of game design, and later, in [38], described an extensive set of 165 detailed heuristics that show exactly how a game could support each condition. They argue that a game that can meet the flow conditions will in general be more likely to induce flow, and therefore more enjoyable to play, than one that does not. However, GameFlow does not encompass an actual method for assessing the occurrence of flow in game players.

The Game Experience Questionnaire, developed by IJsselsteijn et al and first suggested in [15],[1] aims to actually measure flow as well as a number of related (but not mutually exclusive) emotional metrics, including immersion, positive and negative affect, tension, and challenge. For each of the metrics, the questionnaire poses several statements about a game play experience (e.g., for flow, "I was fully occupied with the game"), with possible responses presented in the

[1]At present the questionnaire remains unpublished, although it has been used in a number of studies and can be obtained by contacting the Game Experience Lab at the Eindhoven University of Technology.

format of a five-point Likert-type item. It has since been used in several studies (e.g., [11, 26, 27, 28]) to assess the effects of different game conditions on flow and the other metrics.

Brockmyer et al [4] take a similar approach with their Game Engagement Questionnaire, in which they examine flow, immersion, and presence through several Likert-type items. The questionnaire was designed specifically to assess people's engagement with violent computer games, although the end result is similar in both content and format to the Game Experience Questionnaire. A comparison of the two questionnaires is presented in [30]. We note that *presence*—that is, the feeling of being *in* the fictional game world—may be a useful metric in its own right for game music, especially where the goal of the music system is to provide ambience. An overview of presence as related to other concepts (involvement and immersion), as well as a questionnaire that aims to measure presence in virtual reality environments, is presented in [45].

The applicability of relatively general, experience-based questionnaires (like the Game Experience and Game Engagement questionnaires) to game music evaluation largely hinges on whether multiple musical conditions are feasible and relevant. This is because without multiple conditions there would be no way to discern with certainty whether the music had any effect on the experience, and thus little could be inferred from the results. For example, if a game experience was rated negatively in some way, it could be impossible to tell whether this was due to the music or to some other aspect of the game, such as poor graphics or slow responsiveness.

Instead of testing the effects of different conditions on a player's overall experience, an alternative approach is to place a greater emphasis on the player's subjective opinion of the experience, and how the material being evaluated contributed to it. Paterson et al [31] take this approach in their evaluation of the sound design in a game. They designed a questionnaire that included both open-ended, free-response questions, and more specific Likert scales. The open-ended questions (e.g., "Which part of the game was immersive and why?" purposely did not mention sound so as to not lead the participants to discuss sound if it was not an important factor for them. The Likert scales were more obviously targeted at specific aspects of the sound design (e.g., "The sound made the game feel scary"). This combination was arguably effective because it allowed the participants sufficient freedom to articulate their own opinions about the sound design, while still evaluating more closely the specific items of interest to the researchers. Although their approach could imply a number of relatively bold assumptions—most notably that the participants would actually be able to accurately reflect on their experience—these may not be without some merit, especially if the participants have previous experience playing games. In that case, they may be used to forming preferences about different aspects of games, simply through exposure to different ones. Not surprisingly, research involving game playing typically reports the experience levels of the participants in terms of how often or how long they have been playing games.

This leads to what is perhaps the main limitation of subjective approaches to evaluating player enjoyment, which is that the psychological processes that contribute to enjoyment may operate largely in the subconscious. For example, as previously mentioned, the state of flow is characterized by, among other things, a loss of awareness of oneself [8]. Thus, asking participants to consciously reflect on their experience, possibly of a state of flow, may not yield wholly accurate results. Also, in the case of game music, questionnaires are not conducive to rigorous analyses of small-scale, timed events such as musical transitions or the occurrence of specific chords, for example. This could be a serious limitation since music is by nature a temporal phenomenon, and not only are game music systems inherently dynamic (at least to some extent), but recent game music research has also typically focussed on its dynamic behaviour in particular. Nonetheless, questionnaires, and subjective measures in general, can still offer an important and valid perspective on a player's overall experience.

2.2 Player psychophysiology

The second evaluation question we pose is "Does the music affect the player in the intended way during game play?" In this section, we are concerned with measurable effects that are more objective and specific than overall player enjoyment, although by no means mutually exclusive. For a variety of reasons which we will outline below, we focus primarily on psychophysiological methods, which have been increasingly used in game research over the past decade (an overview is presented in [16]). These methods involve the study of psychological phenomena through the analysis of physiological signals. For example, skin conductance is closely associated with emotional arousal [18, 9], so recordings of players' skin conductance could provide a good, objective estimate of their arousal over time. Common psychophysiological measures in game research include skin conductance, heart rate variability, facial muscle tension, and some others. A comprehensive review of the history and practical use of these and other psychophysiological measures in empirical research is provided in [6].

Kivikangas et al [16] note three main advantages of the psychophysiological approach in the context of game research: First, since the collected data is based on mostly involuntary responses, it is more objective, and not affected by participant bias, limitations of the participant's memory, or limitations of the participant's ability to consciously reflect on potentially subconscious feelings. The latter is of particular importance here since presently it is not actually clear at what level of consciousness game music operates. Second, data can be recorded in real time, without disturbing the player and potentially disrupting a crucial mental state. Finally, psychophysiological measurements are usually sensitive enough to reveal even very subtle responses. We note another important advantage of the psychophysiological approach: it allows responses to be analyzed over time, in relation to specific game events or continuous parameters (including musical ones).

In one of the earliest psychophysiological game studies, Mandryk et al [21] examined correlations between different measurements of game players' psychophysiology and their subjective responses about a game. They carried out two experiments in which participants played a game under different conditions while their physiology was measured (skin conductance, heart rate, respiration amplitude and rate, and facial muscle tension), and then completed a questionnaire that evaluated different aspects of their experience (fun, boredom, challenge, ease, engagement, excitement, and frus-

tration). They found statistically significant correlations between several of the physiological and subjective responses—for example, average skin conductance level was positively correlated with self-reported "fun". This finding was later supported by Tognetti et al [41], who conducted a similar study in which they modelled player preference for different game conditions based on psychophysiological data.

It is perhaps unsurprising that averaged psychophysiological responses could be positively correlated with player enjoyment and preference. A stronger overall psychophysiological response would in general be an indicator of a stronger emotional response, which ostensibly would mean a more enjoyable experience. However, Nacke et al [28] examined the effects of sound (on or off) and music (on or off) in a first-person shooter game on players' psychophysiology and subjective responses, and found no significant correlations with averaged psychophysiological measures, despite finding significant correlations with subjective responses to the Game Experience Questionnaire [15], described above. This suggests that averaging psychophysiological data may be an imperfect or insufficient approach. For example, if a game started with a relatively low level of intensity and then gradually increased in intensity until the end, one might expect the average emotional arousal of the player to be roughly the same as if the intensity was maintained at a more or less moderate level throughout the game. However, the player's experience would likely be quite different, which would probably lead to a clear preference for one version or the other.

The main alternative to the averaged approach is to examine changes in player psychophysiology over time—for example, in relation to specific game events or continuous parameters. Taking the former approach, Ravaja et al [35] used psychophysiological data to analyze the effects of positive (e.g., scoring points) and negative (e.g., losing the game) game events on the player's emotional response. A similar approach could be taken to examine the effects of specific musical events, or of the music in general at designated points in a game narrative—in particular, to see if these effects reflect the game designers' emotional intention.

The notion of an *intended* emotional effect is central to Mirza-babaei et al's [23] idea of Biometric Storyboards. A *Biometric Storyboard* is essentially a set of graphs of psychophysiological data—one of which is manually drawn by a game designer (the intended emotional response), and others that are recorded from one or more players (the actual emotional responses)—over time during game play. The game designer can then review the graphs in order to see any major discrepancies between the intended and actual emotional responses, and tweak the game design accordingly, if desired. The authors showed that using Biometric Storyboards during game development can ultimately lead to game designs that players rate more favourably compared to ones that do not involve user testing.

The occurrence of statistically significant correlations between player psychophysiology and continuous game variables could provide a useful evaluation metric, although to our knowledge this approach has not been explored thus far. It would be relevant particularly in cases where the music is driven by one or more game parameters. For example, in our previous work [34], we describe a game in which the music responds to the narrative by becoming more or less intense and dissonant depending on how close the player is to the enemies, with the ultimate goal being to make the experience more emotionally intense. In this case, the music system could be evaluated in light of this goal by testing for a correlation between player psychophysiology and the intended level of intensity, or by comparing the player's psychophysiology at specific points during the game to a condition with no music.

The use of a psychophysiological method for game music evaluation would generally be underpinned by an assumption that the music should *affect* the player in a particular way. Success would therefore be measured in terms of how strongly or accurately the music does so. It certainly seems reasonable that a game developer would want to make use of a music system that has been shown to affect player psychophysiology in some controllable way. Perhaps the main limitation of approach, however, is that psychophysiological measures can only provide insight into certain known psychological aspects of an experience—they cannot be assumed to represent the overall quality of the experience. However, for these known psychological correlates (e.g., skin conductance's association with emotional arousal; see respective chapters in [6] for other such associations), psychophysiological measures do offer the potential to objectively and continuously examine how music affects the player.

3. MUSIC-ORIENTED APPROACHES

Music-oriented approaches to game music evaluation aim to examine a system's musical capabilities, either on its own or as part of a game. Compared to player-oriented approaches, these have the advantage of not needing the music system to be implemented in an actual game (although this may lead to a reduction of ecological validity). This means that music-oriented approaches may be useful even in early stages of system development, when player-oriented approaches are probably not feasible.

3.1 Aesthetics and style conformity

The third evaluation question we pose is "Does the music sound aesthetically reasonable or stylistically plausible?" Arguably the most basic requirement of a game music system is that it reaches a certain aesthetic standard. Indeed, probably one of the first questions that would come to someone's mind in response to hearing about a particular system is "Does it sound good?" This is particularly relevant in the case of systems that feature *algorithmic music*—music that is generated automatically by algorithms—where "reasonable" aesthetics are generally not taken for granted, at least not to the extent that they could be with a moderately experienced composer. However, even in systems that use human-composed music (e.g., [25]) there is the question of how to effectively transition between different pieces in response to game events, since naively crossfading is prone to harmonic and rhythmic clashing [2]. In this section we review methods that could be used to evaluate aesthetic aspects of game music. We begin by focussing on the evaluation of algorithmic music, since many game music systems that have been proposed thus far use algorithmic music. Furthermore, aesthetic evaluation methods for algorithmic music could, for the most part, be applied to game music as well.

The most common metrics for evaluating the aesthetics of algorithmic music are variations of the Turing Test, first proposed by Alan Turing in 1950 [42]. In the original Turing Test, an "interrogator" uses a text-based interface to inter-

act with two agents, one human and the other a computer program. If the interrogator cannot tell which agent is the human and which is the program, then the program can be said to "think", in Turing's view. In the domain of music, Pearce and Wiggins [33] propose a framework for evaluating algorithmic music systems that uses a variation of the Turing Test. The framework consists of four steps: (1) identify the goals of the system (e.g., to stylistically emulate a particular composer), (2) train the parameters of the system from a relevant corpus, (3) generate music from these parameters, and (4) use the Turing Test variation to compare the generated music against the training music. In their variation, the interrogator listens passively to human- and algorithmically-composed pieces of music (rather than interacting with an agent, as in the original test)—if the interrogator cannot distinguish which is the source music and which is the computer-generated music, then the music system passes the test.[2]

The use of Turing Test variations to evaluate algorithmic music systems has been criticized [1], and their applicability to game music is ostensibly dubious. That is, in general, there is no innate need for an algorithmic game music system to be indistinguishable from human-composed music. However, some element of style conformity may actually be a goal of the system, as it is in [5], for example, where a system is described that is intended to generate music in the style of the Romantic era. In cases like this, a Turing Test variation may very well be applicable, not only to test that the system fulfills the stylistic goal, but also because style conformity would by nature imply that the generated music more or less matches the aesthetic quality of the target music. Indeed, where style conformity excels as an evaluation metric is in its relatively strong objectivity in this regard. Specifically, listeners are not asked for their preference or to think about how "good" the music sounds—instead, aesthetic concerns are, for the most part, reduced to a question of discrimination.

Although style conformity has been the most common evaluation metric used for algorithmic music systems, more broadly, the majority of empirical research into musical aesthetics (see [12] for an overview) has focussed on what leads people to like or otherwise prefer music. As we will soon argue, this need not be a direct concern in game music evaluation, at least in most cases. However, liking and preference could certainly constitute a major design goal of a game music system, in which case (and as with style conformity) there is an implicit assumption that if participants respond favourably (i.e., if they like the music), then the music system has probably reached a sufficient aesthetic standard. Therefore, this could ultimately be an attractive evaluation metric, at least in cases where strong aesthetics are a design goal. Perhaps the two most relevant response formats for recording aesthetic responses are continuous response interfaces (e.g., [20]), in which participants continuously move a slider or dial in real time to indicate their response, and questionnaires. Although a comparison of these two formats is well beyond the scope of this paper, both certainly have their merit.

As we have already suggested, perhaps the most basic requirement of any music system is that it meets a certain aesthetic standard. However, in cases where strong aesthetics are not a driving design goal, it may not make sense to simply ask participants how much they "liked" (or similar) the system's output, as this may lead to a false negative result. Ultimately, at present it is not actually clear whether a game player needs to like or even notice the music at all in order for it to be successful. Of course, we do not suggest not addressing the aesthetic quality of game music—we merely note that there are probably other concerns, such as how well the music conveys the narrative (as described in the following section), and that the aesthetic quality may only need to reach a certain standard in order for the music to be successful overall. A better metric, therefore, may be to instead ask participants if they find that the music sounds "normal", "coherent", or similar, via a carefully designed questionnaire or interview, in such a way that personal preference is minimized.

Evaluating the aesthetics of a game music system is important because, in many ways, the aesthetics could be the weakest link. To state the obvious, if the music was highly unpleasant, it would probably be distracting and therefore lead to a less enjoyable game playing experience. On the other hand, if it sounded pleasant or at least reasonable, then the music would probably be able to function more or less as intended, although it would not necessarily directly lead to a better game. This highlights what is perhaps the main limitation of aesthetic approaches to game music evaluation, which is that they paint a relatively incomplete picture of a system's functionality and intent.

3.2 Conveyance of narrative

The fourth and final evaluation question we pose is "Does the music convey the intended narrative?" As with the question of whether a music system's output affects the game player (see Section 2.2), this question focusses on whether a particular musical function is fulfilled. Specifically, if the music can convey what it is meant to, then it can be considered successful in this regard. Of course, there are many ways to define and think about narrative, but in games there is a tendency for the music to reflect the *emotional narrative*—that is, the changing emotions that coincide with events and changes in game state. For this reason, in this section we focus primarily (but not exclusively) on narrative as a function of emotion. However, the ideas presented here could probably also be applied, to some extent, to research guided by other interpretations of narrative.

There is a large body of research concerned with the relationship between music and emotions. Much of this research has involved participants listening to different pieces of music and rating the emotions they perceive in the music according to a certain model of emotion (see [10] for a comparison of the two main models). Here, we are particularly concerned with the question of whether the emotion that people "hear" in the music (the *perceived emotion*) is actually the same emotion the music was meant to convey (the *intended emotion*). Livingstone et al [19] took this approach in their evaluation of a computational rule system that aims to modify a piece of music in order to express a given emotion, regardless of any emotions present in the original, unmodified piece. They conducted two experiments in which they played emotionally modified pieces of music, and asked participants to identify the emotion they perceived in each one. This allowed the authors to calculate the percentage of cases in which the emotion they intended was "correctly" identi-

[2]Pearce and Wiggins clarify that this does not mean that the music system has achieved intelligence.

fied. Rutherford and Wiggins [36] used a similar approach in their evaluation of a game music system characterized by a single input parameter, *scariness* (a numeric value). They aimed to test whether the *scariness* parameter correlated with the actual amount of scariness, as well as the amount of tension, that people perceived in the music. Accordingly, they created three clips of their system's output, each with a different value for *scariness* (low, medium, and high). They then had participants listen to these clips (amidst others, in order to obfuscate which ones were really being tested), and rate how scary and how tense each one was (as well as other ratings, again, to obfuscate the true intention) on separate bipolar scales. They then analyzed whether the participants tended to rate the medium scariness clip as more scary and tense than the low scariness clip, and whether they rated the high scariness clip as more scary and tense than the medium scariness clip.

This type of methodology could easily be adapted to evaluate a game music system's more dynamic functionality as well, such as how it handles musical transitions. For example, a system could be set to smoothly transition from one emotion to another, and if participants perceive and identify the intended starting and ending emotions, the transition could be considered successful. Continuous response measures (which are outlined in [37]) could alternatively be used for a finer grained analysis of emotional transitions.

The main limitation of comparing intended and perceived emotions is the fact that the identification of emotions is a highly subjective task. For example, the researcher and participant may have different emotional biases, leading to lower-than-expected ratings, which would not necessarily be the "fault" of the game music system. Indeed, it could be that emotions are simply more ambiguous when removed from a more obvious narrative context, such as a game.

Wingstedt et al [44] describe a methodology that could be of direct relevance to the narrative side of game music evaluation, but which does not focus explicitly on emotions. Previously, they had designed a system that allows the user to adjust several features (e.g., tempo, harmonic complexity) of a piece of music in real time via a set of graphical sliders. In their study, they showed participants three videos depicting still scenes—a dark urban area, a view of space from the inside of a spaceship, and a picnic by a lake—and for each one had the participants adjust the sliders in order to make the music best fit the scene. Although the authors were more interested in learning about the participants' understanding of music and its narrative functions, a similar approach could be adopted to evaluate a game music system. For example, participants could adjust the system's parameters so as to best fit different narrative scenes, and then rate according to some scale how satisfied they were with the final results. This approach has two main advantages: First, as we have already suggested, it is conducive to a broader interpretation of narrative, which is not necessarily characterized only by emotion. Second, it allows the system to be evaluated without the researcher having to choose system parameters in order to convey an intended emotion (or other narrative construct). This is instead left to the participants, with the main concern being not so much which parameter values they choose, but how satisfied they are with the music system and its capabilities. However, this may warrant a need for the participants to have some sort of musical background. Additionally, it might not be practical in cases where the music system's interface is relatively complex or has a steep learning curve.

A final approach we will mention that again focusses on narrative in general comes from previous film music studies, in which researchers tested the effects of different musical conditions on viewers' interpretation of ambiguous film scenes [22, 40, 14, 3]. For example, Bravo [3] created two chord progressions that were meant to be similar except in their amount of harmonic dissonance; he then played them in separate conditions alongside the same film clip (an ambiguous scene with a single character), and asked participants to respond to a series of questions about the film. He found that the music affected many of the responses, including their interpretation of the character's emotion, the mood of the scene, and even the genre of the film. A similar approach could be used in order to determine whether different configurations of a music system affect participants' interpretation of a film or game scene in the intended way.

The main advantage of evaluating a music system's ability to convey an intended narrative—emotional or otherwise—is that, if successful, it would demonstrate an element of narrative control, which would certainly be an attractive game design feature. As with player psychophysiology, it makes sense that a game designer would want to be able to control this to some extent. Unlike with player psychophysiology, however, these approaches are relatively subjective and prone to personal bias, particularly when the focus is on emotion.

4. CONCLUSION

Evaluation is an important step in the development of game music systems, but for the most part it has thus far been overlooked. Of course, since game music systems are developed with different design goals and for different purposes, a single, all-encompassing evaluation methodology is probably unrealistic. We have therefore presented four broad methodological approaches for the evaluation of different aspects of a game music system, which arise from two different binary distinctions. We first distinguish player-oriented approaches, which focus on player experience and how the music contributes to it, from music-oriented approaches, which focus on aspects of the music itself. For each of these, we further distinguish between perspectives that are more subjective and aesthetically motivated, and ones that are more objective and functionally motivated. In the former we are concerned with the overall aesthetic quality of the game or music, whereas in the latter we are concerned with whether the music can successfully do what it is intended to do.

The approaches we have presented cover a broad spectrum of evaluation points, although others may certainly be relevant, as we have already suggested. In practice, of course, a well-designed study could probably draw from multiple approaches—for example, it could comprise both psychophysiological measurements during game play, and subjective questionnaires afterwards. In any case, however, we emphasize that evaluation can only increase the value of game music research, as well as provide clear paths to improvement in future research, not only for the music system in question but also for game music as a whole. Indeed, as more is revealed about what does and does not work in game music, different approaches can build on each other, and real progress can be made.

5. REFERENCES

[1] C. Ariza. The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems. *Computer Music Journal*, 33(2):48–70, 2009.

[2] A. Berndt. Musical Nonlinearity in Interactive Narrative Environments. In G. Scavone, V. Verfaille, and A. da Silva, editors, *Proceedings of the International Computer Music Conference (ICMC 2009)*, pages 355–358, Montreal, Canada, 2009.

[3] F. Bravo. The Influence of Music on the Emotional Interpretation of Visual Contexts: Designing Interactive Multimedia Tools for Psychological Research. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval*, pages 600–610, 2012.

[4] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, and J. N. Pidruzny. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634, July 2009.

[5] D. Brown. Mezzo: An Adaptive, Real-Time Composition Program for Game Soundtracks. In *Proceedings of the Eighth Artificial Intelligence and Interactive Digital Entertainment Conference*, pages 68–72, Palo Alto, California, 2012.

[6] J. Cacioppo, L. G. Tassinary, and G. G. Bernston. *Handbook of Psychophysiology*. Cambridge University Press, Cambridge, 2007.

[7] P. Casella and A. Paiva. MAgentA: an Architecture for Real Time Automatic Composition of Background Music. In *Proceedings of the Third International Workshop on Intelligent Virtual Agents*, pages 224–232, 2001.

[8] M. Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper Perennial, New York, 1990.

[9] M. E. Dawson, A. M. Schell, and D. L. Filion. The Electrodermal System. In J. Cacioppo, L. G. Tassinary, and G. G. Berntson, editors, *Handbook of Psychophysiology*, pages 159–181. Cambridge University Press, Cambridge, 2007.

[10] T. Eerola and J. K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[11] B. J. Gajadhar, Y. A. W. de Kort, and W. A. IJsselsteijn. Shared Fun Is Doubled Fun: Player Enjoyment as a Function of Social Setting. In *Proceedings of the Second International Conference on Fun and Games*, pages 106–117, Eindhoven, The Netherlands, 2008.

[12] D. J. Hargreaves and A. C. North. Experimental aesthetics and liking for music. In P. N. Juslin and J. A. Sloboda, editors, *Handbook of Music and Emotion: Theory, Research, Applications*, pages 515–546. Oxford University Press, Oxford, England, 2010.

[13] M. Hoeberechts, R. J. Demopoulos, and M. Katchabaw. A Flexible Music Composition Engine. In *Proceedings of the 2nd Audio Mostly Conference*, pages 52–57, Röntgenbau, Ilmenau, Germany, 2007.

[14] B. Hoeckner, E. W. Wyatt, J. Decety, and H. Nusbaum. Film Music Influences How Viewers Relate to Movie Characters. *Psychology of Aesthetics, Creativity, and the Arts*, 5(2):146–153, 2011.

[15] W. IJsselsteijn, Y. de Kort, K. Poels, A. Jurgelionis, and F. Bellotti. Characterising and Measuring User Experiences in Digital Games. In *Proceedings of 4th International Conference on Advances in Computer Entertainment Technology (ACE '07)*, Salzburg, Austria, 2007.

[16] J. M. Kivikangas, G. Chanel, B. Cowley, I. Ekman, M. Salminen, S. Järvalä, and N. Ravaja. A review of the use of psychophysiological methods in game research. *Journal of Gaming and Virtual Worlds*, 3(3):181–199, 2011.

[17] C. Klimmt, T. Hartmann, and A. Frey. Effectance and Control as Determinants of Video Game Enjoyment. *CyberPsychology & Behavior*, 10(6):845–7, 2007.

[18] P. J. Lang. The Emotion Probe: Studies of Motivation and Attention. *American Psychologist*, 50(5):372–385, 1995.

[19] S. R. Livingstone, R. Muhlberger, A. R. Brown, and W. F. Thompson. Changing Musical Emotion: A Computational Rule System for Modifying Score and Performance. *Computer Music Journal*, 34(1):41–64, 2010.

[20] C. K. Madsen, R. V. Brittin, and D. A. Capperella-Sheldon. An Empirical Method for Measuring the Aesthetic Experience to Music. *Journal of Research in Music Education*, 41(1):57–69, 1993.

[21] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, 25(2):141–158, 2006.

[22] S. K. Marshall and A. J. Cohen. Effects of Musical Soundtracks on Attitudes toward Animated Geometric Figures. *Music Perception*, 6(1):95–112, 1988.

[23] P. Mirza-babaei, L. E. Nacke, J. Gregory, N. Collins, and G. Fitzpatrick. How Does It Play Better? Exploring User Testing and Biometric Storyboards in Games User Research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1499–1508, 2013.

[24] G. B. Moneta. On the Measurement and Conceptualization of Flow. In S. Engeser, editor, *Advances in Flow Research*, pages 23–50. Springer, New York, NY, 2012.

[25] M. Müller and J. Driedger. Data-Driven Sound Track Generation. In M. Müller, M. Goto, and M. Schedl, editors, *Multimodal Music Processing*, pages 175–194. Dagstuhl Publishing, Saarbrücken/Wadern, Germany, 2012.

[26] L. Nacke and C. Lindley. Boredom, Immersion, Flow: A Pilot Study Investigating Player Experience. In *Proceedings of the IADIS Gaming 2008: Design for Engaging Experience and Social Interaction*, pages 103–107, Amsterdam, The Netherlands, 2008. IADIS Press.

[27] L. Nacke and C. A. Lindley. Flow and Immersion in First-Person Shooters: Measuring the player's gameplay experience. In *Proceedings of the 2008 Conference on Future Play*, pages 81–88, 2008.

[28] L. E. Nacke, M. N. Grimshaw, and C. A. Lindley.

More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game. *Interacting with Computers*, 22(5):336–343, 2010.

[29] G. Nierhaus. *Algorithmic Music: Paradigms of Automatic Music Generation*. Springer-Verlag/Wien, New York, 2009.

[30] K. L. Norman. GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interacting with Computers*, 25(4):278–283, 2013.

[31] N. Paterson, K. Naliuka, S. K. Jensen, T. Carrigy, M. Haahr, and F. Conway. Design, Implementation and Evaluation of Audio for a Location Aware Augmented Reality Game. In *Proceedings of the 3rd International Conference on Fun and Games*, pages 149–156, New York, USA, 2010. ACM Press.

[32] M. Pearce, D. Meredith, and G. Wiggins. Motivations and Methodologies for Automation of the Compositional Process. *Musicae Scientiae*, 6(2):119–147, 2002.

[33] M. Pearce and G. Wiggins. Towards A Framework for the Evaluation of Machine Compositions. In *Proceedings of the 2001 AISB Symposium on AI and Creativity in Arts and Science*, pages 22–32, 2001.

[34] A. Prechtl, R. Laney, A. Willis, and R. Samuels. Algorithmic Music As Intelligent Game Music. In *Proceedings of the 50th Anniversary Convention of the AISB*, London, England, 2014.

[35] N. Ravaja, T. Saari, M. Salminen, J. Laarni, and K. Kallinen. Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation. *Media Psychology*, 8(4):343–367, 2006.

[36] J. Rutherford and G. Wiggins. An Experiment in the Automatic Creation of Music Which Has Specific Emotional Content. In *Proceedings of the Seventh International Conference on Music Perception and Cognition*, Sydney, Australia, 2002.

[37] E. Schubert. Continuous self-report methods. In P. N. Juslin and J. A. Sloboda, editors, *Handbook of Music and Emotion: Theory, Research, Applications*, pages 223–254. Oxford University Press, Oxford, England, 2010.

[38] P. Sweetser, D. Johnson, and P. Wyeth. Revisiting the GameFlow Model with Detailed Heuristics. *Journal of Creative Technologies*, 3, 2012.

[39] P. Sweetser and P. Wyeth. GameFlow: A Model for Evaluating Player Enjoyment in Games. *ACM Computers in Entertainment*, 3(3):1–24, 2005.

[40] S.-L. Tan, M. P. Spackman, and M. A. Bezdek. Viewers' Interpretations of Film Characters' Emotions: Effects of Presenting Film Music Before or After a Character is Shown. *Music Perception*, 25(2):135–152, 2007.

[41] S. Tognetti, M. Garbarino, A. Bonarini, and M. Matteucci. Modeling enjoyment preference from physiological responses in a car racing game. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games*, pages 321–328, Copenhagen, Denmark, Aug. 2010.

[42] A. M. Turing. Computing Machinery and Intelligence. *Mind*, 59(236):433–460, 1950.

[43] D. Weibel, B. Wissmath, S. Habegger, Y. Steiner, and R. Groner. Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment. *Computers in Human Behavior*, 24(5):2274–2291, 2008.

[44] J. Wingstedt, S. Brandström, and J. Berg. Young adolescents' usage of narrative functions of media music by manipulation of musical expression. *Psychology of Music*, 36(2):193–214, 2008.

[45] B. G. Witmer and M. J. Singer. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence*, 7(3):225–240, 1998.