Evaluation of musical creativity and musical metacreation systems

KAT AGRES, Computational Creativity Lab, Queen Mary University of London
JAMIE FORTH, Computational Creativity Lab, Queen Mary University of London
GERAINT A. WIGGINS, Computational Creativity Lab, Queen Mary University of London

The field of computational creativity, including musical metacreation, strives to develop artificial systems that are capable of demonstrating creative behavior or producing creative artefacts. But the claim of creativity is often assessed, subjectively only on the part of the researcher, and not objectively at all. This paper provides theoretical motivation for more systematic evaluation of musical metacreation and computationally creative systems, and presents an overview of current methods used to assess human and machine creativity that may be adapted for this purpose. In order to highlight the need for a varied set of evaluation tools, a distinction is drawn between three types of creative systems: those which are purely generative; those which contain internal or external feedback; and those which are capable of reflection and self-reflection. To address the evaluation of each of these aspects, concrete examples of methods and techniques are suggested to help researchers 1) evaluate their systems' creative process and generated artefacts, and test their impact on the perceptual, cognitive, and affective states of the audience, and 2) build mechanisms for reflection into the creative system, including models of human perception and cognition, to endow creative systems with internal evaluative mechanisms to drive self-reflective processes. The first type of evaluation can be considered external to the creative system, and may be employed by the researcher to both better understand the efficacy of their system and its impact, and to incorporate feedback into the system. Here, we take the stance that understanding human creativity can lend insight to computational approaches, and knowledge of how humans perceive creative systems and their output can be incorporated into artificial agents as feedback to provide a sense of how a creation will impact the audience. The second type centers around internal evaluation, in which the system is able to reason about its own behavior and generated output. We argue that creative behavior cannot occur without feedback and reflection by the creative/metacreative system itself. More rigorous empirical testing will allow computational and metacreative systems to become more creative by definition, and can be used to demonstrate the impact and novelty of particular approaches.

Categories and Subject Descriptors: H.5.5 [**Sound and Music Computing**]: Methodologies and techniques

General Terms: musical metacreation, empirical evaluation, computational creativity

Additional Key Words and Phrases: music perception and cognition, artificial intelligence

ACM Reference Format:
Kat Agres, Jamie Forth, and Geraint Wiggins, 2015. Evaluation of musical metacreation and musical creativity. *ACM Comput. in Ent.* 100, 10, Article 1 (May 2015), 35 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

We live in a period of significant advancement in the area of music creation, with new approaches, techniques, and theoretical dialogue surfacing in a broad range of research circles. Rapid growth has recently been seen in musical metacreation (MuMe)

Author's addresses: K. Agres, J. Forth, and G. Wiggins, School of Electronic Engineering and Computer Science, Cognitive Science Group, Queen Mary University of London, Mile End Road, London E1 4FZ, UK.

DOI: http://dx.doi.org/10.1145/0000000.0000000

in particular, an area within the field of computational creativity (CC). Colton and Wiggins [2012, §1] provide a working definition of CC as

> "the philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative."

Research within MuMe specifically seeks to automate aspects of musical creativity, with the aim of producing systems or artefacts that would be deemed creative by unbiased observers. Echoing the above definition of CC, Pasquier et al. [2014] consider a computer program to be a metacreation "if it exhibits behaviors that would be considered creative if performed by humans." In spite of recent growth within MuMe, however, there is commonly little systematic evaluation of metacreative systems or artefacts, especially in empirical terms. As MuMe is considered to be a branch of CC, research in this area should adhere to the empirical and evaluative standards established by the larger CC community (e.g., Jordanous [2012]), which in turn will help to promote the advancement of both fields.

Considering recent developments in MuMe, it is insightful to highlight the difference between generative models and creative models of music (as discussed, for example, in the Proceedings of the 13th International Conference on Computational Creativity; see also Wiggins et al. [2009]). Although many generative models have become quite sophisticated, they do not contain an element of reflection or evaluation, and therefore might not necessarily be considered creative systems in their own right (see Figure 1). For example, it is not uncommon for generative systems to rely on the human designer and/or user to preside over matters of aesthetics, which results in the lack of an inbuilt capacity for evaluation or self-reflection. Naturally, the development of generative systems may stem from motivations other than the investigation and implementation of fully creative systems. However, the consideration of metacreation systems from a perspective of CC may present previously unconsidered theoretical and practical opportunities, including music generation based on a model of how the audience is likely to respond.

In an attempt to produce a set of meaningful, quantitative and qualitative measures of musical creativity, one may find clues by examining how humans exhibit creative musical behavior. Individual human creativity, and self-evaluation of created artefacts, are dependent upon the individual's domain-specific expertise, the social experiences and social climate in which the creator is embedded, and the individual's particular goals at hand [Wiggins et al. 2015]. The specific audience at hand is also a key factor, especially within the context of computationally-generated music, as academic and domain experts may focus on different aims or attributes of the systems and their output [Eigenfeldt et al. 2012]. By considering the processes underlying creative behavior and self-evaluation in humans, we can learn how to eventually build these techniques for reflection and critique into musical generative systems. Incorporating humanistic aspects of the creative process are not a requirement for MuMe systems, of course, and computational processes are likely to differ from those of humans; however, knowledge of human perception and production may provide helpful insight for achieving the MuMe system's goals. In addition, findings about the ways in which humans judge artificial systems are also discussed, because biases and expectations relating to artificial systems and technology can greatly influence creative evaluation.

To identify mechanisms underlying musical creativity and evaluation in humans, we consider a range of empirical findings from different areas, including behavioral psychology (e.g., lateral and divergent thinking tasks), cognitive science (e.g., learning and expectation mechanisms underlying knowledge representation and creative behaviors), relevant theoretical constructs (such as conceptual spaces theory), and neuro-

science (e.g., examining brain response whilst engaging in creative behavior to assess the neural processes and functional organization of brain areas implicated in creative tasks). After examining relevant behavioral and cognitive findings as outlined above, we will discuss how methods of feedback and self-evaluation may be implemented to extend the creative potential of generative models. We posit that the grand challenge of this area is to model what the viewer/listener will perceive and how she will react, which requires a listener model, reflection, and dynamic response to this feedback and reflection. The range of methods outlined in this paper is by no means exhaustive, but we hope to support an expansion of the current methods used for testing computational creativity (as well as the theoretical motivation for more rigorous empirical evaluation), and extend our conclusions to the area of MuMe in particular.

The structure of this paper reflects our focus not only on the evaluation of generated artefacts and underlying creative processes, but also on the evaluation requirements of and for systems including feedback and reflection.

In order to suggest evaluation methods for these different purposes, we consider three important aspects of creative systems, as illustrated in Figure 1: first is the purely generative part, in which a software system, via some generative process, produces artefacts which are then presented to an audience. The second aspect is feedback, which may consist of the creative system's own processes or generated output (for example, a MuMe system may take into consideration the previously generated melody to compose its next melody), or contain direct or indirect psychological, cognitive, and/or affective responses from the audience. Finally, the third aspect is self-evaluation through reflection, which can again be based on the system's own processes and output, as well as on reasoning about the impact of these on the listener or audience.

The first two aspects can be addressed by implementing evaluation methods which are external to the system, to assess the creative process, generated artefacts, and audience perceptual/cognitive/affective states. In addition to testing the system and its impact on the audience, these methods may be used to build feedback into the system, to influence real-time creative processes. The third aspect, reflection, may be addressed by incorporating a model of the system within the context of its performance space, or a model of some aspect of the audience. Accordingly, Section 4 discusses two areas of external evaluation: methods to evaluate characteristics of the creative process and generated artefacts, and methods to test the audience's perceptual and emotional responses to the system and its output. Then, Section 5 focuses on internal evaluation by providing examples of techniques which may be used to model audience states, which are useful if not necessary for building reflection into a creative system. Before delving into these methods, we first contextualize our discussions by providing an overview of theoretical approaches to creativity (Section 2), and frame the importance of evaluation in relation to common goals of MuMe and CC systems (Section 3).

2. THEORETICAL PERSPECTIVES ON CREATIVITY

An early attempt to formalize the investigation of creativity was by Arthur Koestler [1964], who formulated a general theory of human creativity that introduced the idea of *bisociation*: a blending of ideas drawn from previously unrelated semantic matrices of thought. The theory focuses on abstraction, categorization, and analogy as a means of forming bisociations. This approach was very influential for subsequent models of creativity, especially the theory of *conceptual blending* [Fauconnier and Turner 2008; Tunner and Fauconnier 1995]. Conceptual blending provides a framework and terminology for discussing creative thought processes, and for outlining how existing concepts may be combined to form a novel blend of ideas. It is not difficult to imagine

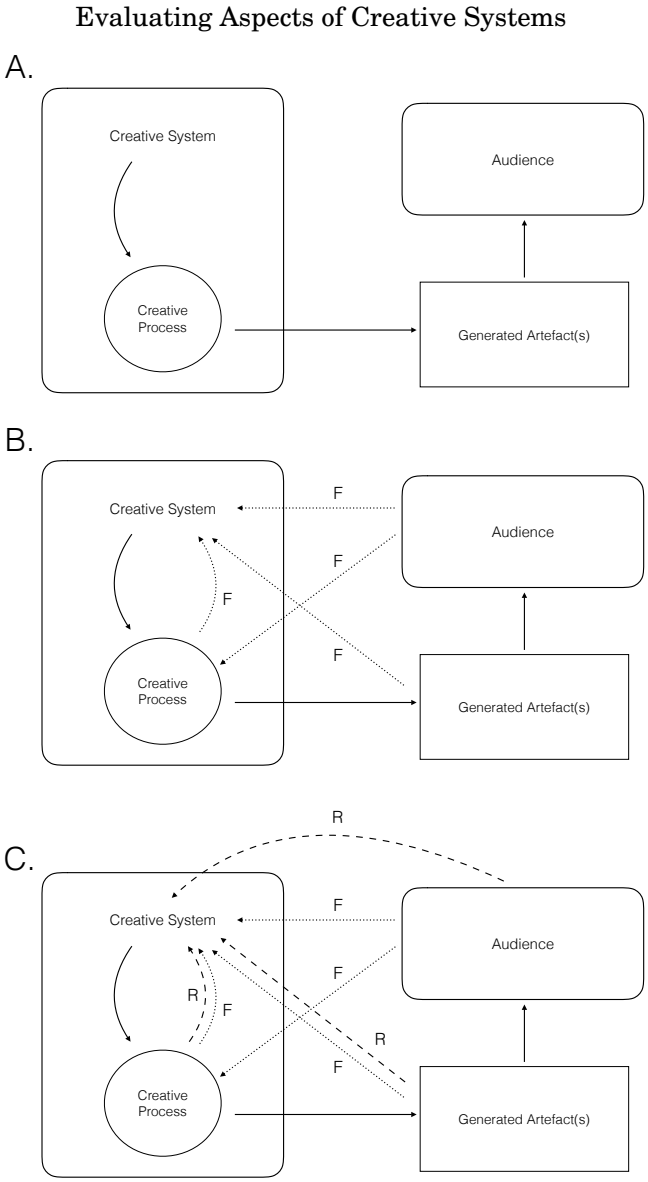

Fig. 1. Illustrating the different aspects of creative systems which may benefit from formal evaluation: The systems range from the merely generative (top), to the fully creative, reflective system (bottom). **1A.** The solid lines from the Creative System to the Generated artefact and ultimately to the Audience represent the flow of information through the process of generation without feedback or reflection. **1B.** In this system, the dotted lines marked with an 'F' represent feedback from the Audience, or from the Creative Process or Generated artefacts themselves. **1C.** This sophisticated type of system incorporates reflection, as denoted by dashed lines marked with an 'R': information regarding the Creative Process, Generated artefacts, and/or the Audience is taken into consideration by the creative system for self-assessment and reflection. It is this last element of reflection that is necessary for qualification as a truly creative system. Note that it is technically possible for a reflective system to not contain feedback, but because this scenario is rare, we focus on the three distinctions illustrated above (purely generative systems, systems containing feedback, and systems containing reflection). Note also that the context in which the Audience and Creative System are embedded is not explicitly shown, but taken to be an implicit part of the Creative System and Audience.

how this approach is useful in a computational setting, for developing new melodic, harmonic, or rhythmic content based on existing material.

In her seminal book *The Creative Mind*, Margaret Boden [2004] identifies three different types of creativity, relative to the notion of the *conceptual space* which contains all the concepts of a particular kind: *combinatorial*, *exploratory*, and *transformational*. Combinatorial creativity, similar in principle to the conceptual blending framework mentioned above, is the process of amalgamating elements of existing concepts to form a new concept, and, where the concepts combined are from different conceptual spaces, forming a new conceptual space as well. For example, within a particular conceptual space of musical genres, one can discover novel combinations of features to meet an artistic goal, e.g., combining a traditional folk melody with blues harmonization, or generating idiomatic jazz variations of melodies by J. S. Bach as demonstrated by Tzimeas and Mangina [2006]. Exploratory creativity involves discovering novel concepts or items within an existing conceptual space. For example, within the existing rhythmic schemata and syncopation structures of Latin jazz or traditional West African music [Tzimeas and Mangina 2009], discovering a previously unencountered rhythmic pattern within the space. And lastly, transformational creativity involves changing the structure or rules defining the conceptual space itself, or the type of traversal through this space. Transformational creativity thus produces a shift in thought or paradigm, enabling a new set of possible created artefacts. Modern technologies that alter the types of sounds we are able to produce or enable new kinds of creative interaction between human performers and artificial systems are candidates for this type of boundary-shifting creativity [Wiggins and Forth 2016]. Boden also offers the distinction between *psychological creativity*, or *P-creativity*, which refers to a particular individual's creativity, and *historical creativity*, or *H-creativity*, which is creativity recognized as novel at the societal level [Boden 2004].

Geraint Wiggins has formalized Boden's conceptual spaces theory in his creative systems framework (CSF) [Wiggins 2006a,b]. The CSF is a theoretical framework for describing and comparing creative systems, conceptualized as computational processes, but not excluding human components [Wiggins and Forth 2016]. Thus it provides a shared terminology which enables discourse about creative systems [see, for example, Ritchie 2007; Maher 2010]. The CSF formalizes Boden's notion of exploratory creativity as a well-defined process of traversal through the universe of partial concepts. This search procedure takes into account a rule set specifying a subset of the universe, corresponding to Boden's idea of conceptual space. The system must be able to internally evaluate the artefacts it encounters, which may include discoveries that do not conform to the rules specifying the current conceptual space—so-called *aberrant* concepts—that may or may not be valuable. The discovery of aberrant concepts may require that a system revises its conceptual space and/or traversal mechanism. Boden's proposal of transformational creativity is formalized in the CSF as exploratory creativity at the meta-level, where the conceptual space is the conceptual space of conceptual spaces, rather than a conceptual space of concepts or artefacts. Understanding this relationship between exploratory and transformational creativity leads to the conclusion that transformational creativity necessarily involves reflection: the system must be capable of reasoning about itself, either in response to external feedback, or with respect to internal evaluative mechanisms. By describing MuMe systems in terms of the CSF, the designer can relate aspects of a specific implementation to wider theories of creativity. Furthermore, the formalism facilitates precise specification of hypotheses and appropriate evaluation criteria, and importantly enables results to be comparable between different systems.

Boden [1998] raises the issue of evaluation as a fundamental aspect of creativity. We note that there are two aspects of evaluation in creative systems: first, the eval-

uation of the system itself, viewed from the outside, as a scientific and/or creative contribution, which may concern either the creative outputs, or the creative process, or both; and, second, the evaluation by the system itself of its own outputs, as part of the creative process. In the former case, evaluation may be implicit, in, for example, a heuristic search process, and so identifying it may require subtle analysis. Again, the CSF provides tools for thinking about evaluation within a creative system.

Wiggins et al. [2015] formulate evaluation as a function of four parameters: artefact, creator, audience, and context. This gives a clear signal of how difficult the problem is in general. However, for the purposes of MuMe, these parameters can be pinned down: the creator is generally a hybrid system including the music-programmer and her software; and the audience and context are also well-understood. Thus, the problem becomes more manageable, as illustrated by Wiggins and Forth [2016].

3. CONTEXTUALIZING EVALUATION: AREAS OF INTEREST FOR MUME AND MUSICAL CC SYSTEMS

Now that we have considered theoretical perspectives for examining creativity, and motivated the need for different types of evaluation depending on the type of MuMe system (generative, generative with feedback, or fully reflective), it is helpful to ground our discussion in practical application. We therefore present some of the main areas of interest to computational creativity in music and MuMe in order to provide tangible applications of evaluation methods that may be applied in these areas. Below we discuss common goals of these systems, including the generation of melodic, harmonic, and rhythmic content, creating an affective response in the audience, and developing interactive systems that take the performer or audience into consideration during creative processing. The techniques we discuss here are applicable or relevant to each of these different levels and aspects of music creation.

3.1. Melodic and Harmonic generation

The problem of generating melody and/or harmony computationally is one of the oldest areas of interest in computer-generated music, probably because the earliest computers were not capable of generating audio for themselves, so they were conceptualized as score-generators. The earliest known computer-generated score was the *Illiac Suite* [Hiller and Isaacson 1959, 1993], for string quartet, in which musical textures were generated statistically. The *Illiac Suite*, which was named after the computer used to generate it, was presented purely as a work of art (though the methods used were explained). The purely artistic perspective meant that evaluation, in the scientific sense, was unnecessary: the work is what it is, and no formal evaluation will change that.

Other notable pieces of work in this area are by Cope [1987, 1991, 2005] and Ebcioğlu [1988, 1990]. Cope uses a template based approach[1] in which the harmonic and melodic content of one human-composed work is adapted into the rhythmic structure of another. It is evaluated by means of a so-called "musical Turing Test"[2] [Cope 2005] where humans are invited to approve of it as stylistic pastiche at a public concert. While this may give a general stamp of approval, the approach is fraught with difficulty, because it is very broad-brush, so it does not give specific information with which to improve systems or theory, and, since there is evidence that humans may be biased against computers in creative contexts [Moffat and Kelly 2006], the results may be skewed.

[1]Cope does not explain his work in reproducible detail, so this claim is based on a report by Douglas Hofstadter in the same book [Wiggins 2007].

[2]In fact, the nature of the Imitation Game, as specified by Turing [1950], is rather different from this evaluation. Ariza [2009] provides a comprehensive analysis of Turing-like and other discriminatory tests in the context of generative music.

Ebcioğlu's CHORAL system is an archetype of "Good Old-Fashioned AI" in which a highly detailed rule system is combined with a specially designed intelligent backtracking language designed to deal with the highly problematic search space of harmony (in this case, in the style of J. S. Bach). Phon-Amnuaisuk et al. [2006] and Hörnel and Menzel [1998] also focused on Baroque harmony, the former echoing Ebcioğlu [1988] in focusing on search control, and the latter employing then cutting-edge artificial neural network technology. Part of the reason that J. S. Bach is such a strong focus of interest is that his style is copiously studied and explicated in the musicology literature, so that harmony generated in this style is relatively easy to evaluate, as least as a stylistic pastiche generator, because a good model exists with which to compare it. This is not the case for the vast majority of musical styles. Pearce and Wiggins [2007] were more unusual in considering the generation of melody in Baroque Chorale style. Again, good music-theoretic understanding of the genre made evaluation and—uniquely in this literature—resulting specific improvement of the system possible, by means of a version of the Consensual Assessment Technique [Amabile 1996b], discussed below in section 4.2. More recently, Brown et al. [2015] have developed a number of melodic generation algorithms based on preference rule and statistical models of music cognition.

Another well-theorized area is jazz, in both melody and harmony. Numerous researchers have worked on jazz melody and harmony, and, in the case of the latter, there is a good model on which to work [e.g., Pachet 1990; Biles 1994; Steedman 1996; Papadopoulos and Wiggins 1998; Lewis 2000; Pachet 2002].

A successful researcher with a sightly different aim is Robert Rowe [2001]. Rowe's *Cypher* system, begun in the late 1980s, and still being used, is a performance environment specifically intended for live computational generation of music from a "score", but driven by instrumental performance, connected to the computer via MIDI. In this work, the harmonic, melodic and rhythmic structures generated by the computer are defined by a human composer, and produced in response to stimulus from a human performer. As such, it is rather difficult to draw a clear line around generation by the machine itself, since the overall effect is very much a hybrid of human (composer and performer) and computer.

Notwithstanding the exceptions of Jazz and Bach, where a strong music theory provides a basis for semi-objective evaluation on the basis of style, there is a general dearth of strong models against which to compare musical outputs, because to do so completely would entail solving numerous open problems of music perception, and/or involve substantially more detailed music-theoretic accounts than are normally available: Chuan and Chew [2011] evaluate their harmonic sequence generator by reference to the harmonic analyzer of Temperley and Sleator [1999] and in terms of information-theoretic measures, but this is the exception and not the rule. Another noteworthy example is the harmonic sequence generator of Burnett et al. [2012], which also provides a useful discourse on the evaluation of MuMe creativity (by re-framing the question concerning the extent to which a computer system is creative) in terms of how distinguishable the computer-generations are from human-generated examples. The generation of electronic dance music is also particularly interesting in this regard as the style itself is the result of heavily automated production techniques, and offers the potential for evaluation in terms of contemporary cultural practices, which may be probed in the form of listening tests, on-line questionnaires, or even by covertly releasing computer generated compositions into the public domain in order to elicit qualitative and quantitative feedback [e.g. Collins 2008; Anderson et al. 2013]. However in general, where comparators do exist they pertain only to specific stylistic pastiche, and then in a more-or-less informal way [e.g., Biles 1994; Phon-Amnuaisuk et al. 1999; Cope 2005]. This places the *scientific* study of MuMe, beyond the purely technological, in a

difficult position. Our solution is to look beyond the musicological and stylistic, into the psychological theory that underpins hearing and therefore music cognition [Wiggins 2009, 2012a,c,d]. Thus, we claim, we may be able to find general approaches which help us to evaluate MuMe, not only in the context of stylistic generation, but also in more generally creative terms.

Because of the history of music education, one method of evaluation for musical harmony is rather generally applicable, where the aim is a given style. It is common in music courses to require students to harmonise melodies in particular harmonic styles, that of J. S. Bach being a frequent example. This means that a music educator may be asked to assess the outputs of a system, in terms comparable with the assessment of students. Thus, a comparison between computer and human creativity can be achieved, though in a context which is limited to the solving of set problems [Phon-Amnuaisuk et al. 1999]. In principle, of course, the same idea can be used with all aspects of creativity, but with the same caveat.

3.2. Rhythmic generation

There are two ways of interpreting the epithet of rhythmic generation: the general temporal aspects of musical organization that are experienced as meter and rhythm; and the specific generation of percussive instrumental parts—scores or tracks providing rhythmic accompaniment within a musical texture. The vast majority of extant musical creativity work falls into the former category. Generation within both categories may pertain to compositional structure and/or expressive performance timing. It is also important to note that there is a substantial community interested in the study of psychology and neuroscience of rhythm and meter, with its own website[3]; this literature affords a wealth of opportunity for MuMe practitioners interested in sharing their creative responsibility.

Meter and rhythm are very much open questions in MuMe. While there is a certain artistic potential in the extremes of computer performance (in particular, very fast and/or very precise playing), these effects can only contribute so much in musical terms, and anyway have been the subject of wider musical enquiry for some time, for example, the player piano works of Nancarrow dating back to the late 1940s, or more recently, the advent of the Moog sequencer in the early 1970s.[4] For MuMe to develop beyond these restricted means of expression, more theoretical development is necessary. The problem, perhaps, is that meter and rhythm are so fundamental to human musical behaviour that we do not even notice that we are doing them; this proposal is supported by the success of John Cage's random music, in which it is easy to hear rhythm, even though, because the sounds are random, it is—objectively—not there, a phenomenon known as subjective rhythmization [Bolton 1894].

Music theorists have studied meter and rhythm extensively, however. One recent and successful theory is that of London [2012], which has been formalized as a *conceptual space* [Gärdenfors 2000] by Forth [2012]. The premise of London's view is that meter is a cognitive temporal framework within which rhythm takes place. The same ideas, in restricted form, are encoded in Western music notation. Even if one does not

[3]http://rppw.org

[4]The ground-breaking Tangerine Dream album, *Phaedra* (1974), opens with the transformation of a sequence played so fast that it sounds like a complex timbre in its own right, which then slows and drops in pitch until it eventually becomes a bass line, played with inhuman regularity. In this one short musical section, Tangerine Dream both brought to the foreground in a popular music context deep questions about the perception of timbre and musical structure—echoing the earlier music-theoretical writing of Cowell [1930], and compositional techniques explored by composers such as Stockhausen and Tenney—and set the stage for pop classics throughout the 1970s and 1980s, such as *I feel love* (Donna Summmer and Giorgio Moroder, 1977).

agree with this epistemology, it provides a useful starting point for the algorithmic construction of rhythmic music. Forth's formalization, though complicated, may therefore prove useful in MuMe, affording a mathematical space of meter and rhythm which can be navigated in rigorous ways.

In activities where music is generated to accompany an existing line, be it melody or otherwise, meter is *de facto* pre-defined. The problem of rhythm then reverts (again) to stylistic pastiche: the aim is to produce structures that match the existing part, and whatever instrumental intent is involved. For example, in Ebcioğlu's system, the chorale melody supplies the essential structure, and movement in counterpoint is considered as passing notes. In Pearce and Wiggins' [2007] melody generation work, the melody was generated *given* a rhythm (in literal statistical terms).

There are many examples of generative systems with particular attention to rhythm, although less specifically consider rhythmic creativity in conjunction with psychological aspects of metrical perception. In the field of style replication, Bel and Kippen [1992] and Wright and Wessel [1998] are notable examples. Bel and Kippen [1992] in particular were ahead of their time in implementing a generative system for a particular well-defined musical practice (Tabla drumming) and in evaluating it "in the wild" by seeking appraisal of the musical quality of the work from Indian tabla masters (who approved of a good proportion of the generated music). In discussing the rhythmic aspects of this work, it is important to acknowledge that the method used involved appropriation of a language used for describing tabla drum sequences for didactic purposes. Treating a rhythmic sequence as a sequence of arbitrary symbols brings the work closer to the harmonic and melodic sequence generation systems outlined in the previous section than might be naïvely expected. In an explicitly pedagogical context designed to facilitate effective instrumental practice, Hawryshkewich et al. [2010] describe a Markov model-based system that learns the style of an individual drummer and is able to "play along" in a manner comparable to Pachet's [2006] keyboard-based Continuator system.

Beyond stylistic replication, Kaliakatsos-Papakostas et al. [2012] investigate the qualities of L-systems in the domain of rhythmic generation, producing a modification of the classical formulation, termed *Finite L-systems*, which are generative grammars exhibiting behavior particularly well-suited to the generation of rhythmic structure. This work uses information theoretic and compression rate measures to quantify rhythms generated by both standard and modified L-systems, enabling conclusions to be drawn about the relative strengths of the new grammar form in terms of the control it affords over the complexity of, and repetition within, generated sequences. In combination with a psychological model of time perception, L-systems have also been used for the generation of expressive timing [Patricio and Honing 2014].

The multi-agent system paradigm has also been used to generate rhythmic sequences, typically motivated by the potential of such systems to produce emergent behavior. Levisohn and Pasquier [2008] describe a metacreative system for rhythmic generation comprising six agents with simple predefined rule-based behaviors. The behaviors are defined in accordance with a subsumption architecture, whereby rules are grouped into layers, with different layers taking precedence depending on context. The impact of different rule-sets on the resulting emergent behavior, and consequently the generated rhythmic output, is evaluated in terms of quantified measures of convergence. A hypothesis of this work is that utilizing a subsumption architecture enables agent behavior to be defined by simple, intuitive rules, which are capable of producing emergent and unexpected behavior comparable to systems embodying more complicated behavioral rules. An example of the latter approach, albeit entirely within an artistic context, is taken by Eigenfeldt [2007], which involves more complicated forms of agent behavior including fuzzy logic and a model of societal interaction. A more

music-theoretical approach, but again lacking any form of evaluation, is described by Pachet [2000], whereby agents are equipped with rules encoding basic musical knowledge enabling them to generate, modify and compare musical structures. In a game-like scenario, agents evolve new music together by trying to satisfy the shared high-level goal of producing "coherent music".

3.3. Creating an affective response in the listener

Most authors agree that the primary aim of music composition and performance is to achieve an affective response in the listener. Some go further and suggest that a composer's aim is to align the listener's own brain states with his or her own [Bharucha et al. 2009]. Similarly, one of the main reported reasons for listening to music is to alter or amplify mood [DeNora 2000].

Affect in music is an extremely difficult area, because of its deep subjectivity. One aspect that does seem generally agreed is the human preference for music which sounds as though it has been played by a human (except where inhuman features, such as extreme precision or speed, are part of the aesthetic), and this would be a fruitful area of research for MuMe: a tradition exists, in the RENCON music rendering contest[5], on which to draw. Achieving specific affective response in a human listener is challenging for a human expert composer, let alone the algorithms used in MuMe. Nevertheless, some researchers have attempted to algorithmically generate music mapping onto particular emotions, or music based on manipulable levels of psychological arousal and valence, however many of these cases are in want of empirical evaluation with listeners [Wallis et al. 2008].

As a notable exception to the lack of rigorous testing with humans in this area, some researchers working at the intersection of CC and affective computing, notably Kristine Monteith, Dan Ventura and colleagues, have addressed both the generation and evaluation of computer music intended to elicit specific emotions or induce particular physiological responses [Monteith et al. 2013], such as changes in breathing and heart rate. For example, one system uses statistical distributions from musical corpora representing different emotions to generate music matched with a target emotion [Monteith et al. 2010]. Other work aims to explore how computationally generated soundtracks may guide or enhance perceived emotion in concurrently-spoken stories [Monteith et al. 2011; Rubin and Agrawala 2014]. In regard to affective response, the present paper focuses on how MuMe systems may be used not only to evoke certain emotions in listeners, but how listeners' subjective and physiological responses may be used as feedback for generative systems.

3.4. Interactive systems: Incorporating composer, performer and audience feedback

There are broadly four ways in which the behavior of a musical algorithm can be influenced by human interactors: direct engagement with algorithmic processes at compositional or computational levels, either via code or graphical user interfaces; instrumental control, for example by MIDI or OSC; explicit feedback, for example with a human indicating like or dislike for the music; and by implicit feedback, for example by measuring human physical responses such as heart rate, to determine arousal in response to the music. By considering these human intellectual, emotional, psychological, and physiological factors, mental and physical *states* may be modeled and incorporated into music metacreative systems.

As noted in section 3.1, the use of computers as tools within compositional practice is one of the oldest forms of generative music-making [for a survey, see Assayag 1998].

[5]http://renconmusic.org

The field of computer-assisted composition is primarily focused, although not exclusively, on the manipulation and generation of musical scores. More generally, the term can be said to apply to the processing of symbolic representations of abstract musical and compositional ideas, which may include perceptual or psychologically-motivated techniques. This does not mean that other aspects of computer-based music making, such as sound synthesis or interaction design, are precluded from consideration within the frame of computer-assisted composition, but that the primary concern has traditionally been understood as the unification of instrumental compositional practice and computational methods. Consequently, computer-assisted composition is typically characterized as an off-line mode of interaction between human composer and computer system. Compositional software environments must balance the often conflicting requirements of providing composers with tools and abstractions for representing and processing musical information, while simultaneously offering freedom for customization and the capacity to incorporate new abstractions that may emerge as part of the compositional process. The OpenMusic system, and its predecessor PatchWork, are notable examples in the literature given their long periods of active development and wide adoption among leading contemporary music composers [Assayag et al. 1999]. An early system employing genetic algorithms by Horner and Goldberg [1991] embodies a very direct mode of interaction whereby a composer can specify an initial musical idea and delegate its development to the system. New material is in turn presented to the composer for evaluation and subsequent processing. More recent approaches have focused on developing domain-specific programming languages for music processing that are more aligned with mainstream computer languages and associated technologies, such as SuperCollider3 [McCartney 2002], ChucK [Wang et al. 2015], and Tidal [McLean and Wiggins 2010]. These languages afford composer-programmers greater flexibility in defining and controlling the computational processes involved in compositional thinking, while also enable consideration of interactive and performance dimensions, further eroding the boundaries between composition, improvisation and live performance.

Performer influence has been present in computer music since the 1970s, most notably from IRCAM in Paris, where a succession of composers have supplied parameters to music algorithms from human performance, and in other work such as Robert Rowe's Cypher-based music [Rowe 2001, 2008], Peter Nelson's *Gross Concerto*, Jean-Claude Risset's *Trois études en duo*, François Pachet's Continuator system [Pachet 2006], and many others. The MaxMSP programming language, originally developed at IRCAM, and its successor, PD [Puckette 1996], has enabled a significant amount of performance-driven computer music to be developed. To our knowledge, the vast majority of this work does not attempt to enter the field of computational creativity, but exists as art, not open to scrutiny by scientific evaluation.

The *Cypher* system, mentioned twice above [Rowe 2001, 2008] takes input and gives output in MIDI. It affords performer control possibilities comparable to MaxMSP, though in a more score-based, less general paradigm.

In these systems, and others like them, it is possible to introduce feedback into the music generation algorithms, explicitly, via, for example, on-screen controls, or implicitly, via inputs from appropriate sensing devices. It is clear that the former of these has been achieved, because any kind of parametric control over the algorithms by a composer or technician working the software would fulfill the definition; however, this is not normally (if ever) conceptualized as preference feedback, but rather as straightforward performance control. Perhaps, therefore, MuMe needs to consider how to share creative responsibility in this context [Colton and Wiggins 2012]: instead of controlling the computer, maybe the musician could advise it.

Influencing performance by audience response is more problematic because of the difficulty of gathering the necessary response data. There have been attempts to allow computational performances to react to audience feedback given by coloured cards, for example. However, there is a tendency for such responses to average out across human subjectivity, and therefore little is gained. Stockholm and Pasquier [2008, 2009] partially address this issue within a system designed to enable a collective soundscape performance. The performance involves the playback of pre-recorded music, which is streamed to participants' laptops via a web-based interface. The music is composed to encompass a range of "moods", and the composer provides basic descriptive metadata and classifies each audio file within a two-dimensional valence-arousal space. Audience members are encouraged to share their mood with the system, which will accordingly select an audio track—taking into account other material currently playing in the space—to stream to their laptop. Participants can provide feedback at the end of each section of music as to how accurately the music reflected their mood, which is incorporated by the system using a simple form of reinforcement learning. Evaluation is minimal and anecdotal, of the form: "audio performance is clearly shaped by participant moods" or performances were "successful at inspiring conversations to arise between disparate people in the performance environment" [Stockholm and Pasquier 2008, pp. 565–566]. However, the approach of preserving a degree of audience individuality within the performance does alleviate the problems associated with simple averaging-based approaches to incorporating audience feedback. The McGill University performance lab, led by Stephen McAdams, is an extraordinary resource in a concert hall that allows individual audience feedback, both explicit, via touch interfaces, and implicit, via physiological measures such as heart rate and skin conductance. The lab is intended for music perception research. Such a resource could be used to great effect for artistic purposes—if the extended preparation time required to install the audience were not prohibitive.

Considering the general issue of feedback raises questions of representation and what meanings, in musical terms, can be associated with feedback data. Livingstone et al. [2005] outline an "affective performance framework", which provides a model of listener attitudes and a means of annotating compositions with emotional metadata. The purpose of this work is to enable greater realism and expression in computer performances of music by incorporating knowledge of intended emotional expression; however, a similar approach could be adopted to allow performer or audience feedback to influence generative systems at higher levels of musical conceptualization.

We conclude that, while there are too many existing approaches to instrumental control of algorithms to mention here, other kinds of more subtle human control are mostly absent in MuMe, and where they are present, they are conceptualised as parametric control. In the explicit computational creativity context, there seems to be little work in which the creative responsibility is genuinely shared with the computer.

4. EXTERNAL METHODS: EVALUATING THE CREATIVE PROCESS AND ARTEFACTS BY TESTING PERCEPTUAL, COGNITIVE, AND AFFECTIVE RESPONSES IN THE AUDIENCE

In this section, we begin by discussing methods which address human creativity, as these are relevant and illuminating for CC and MuMe research. After providing this background, specific methodologies and applications that can be applied to the study of MuMe will be discussed, making reference to the common areas of interest introduced in Section 3. Because MuMe systems have historically relied heavily upon the researcher's own subjective opinions for evaluation of these systems (which is neither rigorous nor replicable), we place emphasis on quantitative rather than qualitative approaches. It is not possible to list *all* possible methodologies which may be of use; the following highlighted examples are offered as a selection of some of the most useful

Table I. Overview of evaluation methods for creative systems

*Methods of **external evaluation**: Evaluating the creative process and artefacts by testing perceptual, cognitive, and affective responses in the audience*	
§4.1	Behavioral tests: adaptations of convergent and divergent thinking tests for evaluating the creative process and artefacts
§4.2	Consensual Assessment Technique: evaluation by experts
§4.3	Evaluation criteria from computational creativity
§4.4	Questionnaires, correlational studies, and rating scales as a means of 1) directly evaluating features of the creative system, and 2) indirectly testing the system by correlating perceptual, cognitive, or affective responses with features of the system)
§4.5	Physiological measurements (e.g., reaction time; movement and motion capture; eye-tracking)
§4.6	Neurophysiological measurements (e.g., EEG)
*Methods of **internal evaluation**: Modeling human creative behavior and audience perceptual, cognitive, and affective states for self-reflection*	
§5.1	Self-assessment based on behavioral tests: the system reflects upon its own behavior)
§5.2	Reflective models based on external tests of audience perception, cognition, and affective states
§5.3	Measures of prediction and expectation for modeling perception and affective response in listeners
§5.4	Conceptual representations for self-reflection

methods for the evaluation of MuMe and musical CC systems. For further information on qualitative and quantitative methods and evaluation techniques, we recommend consulting a reference such as Mertens [2014].

It is worth mentioning here that some recent research has made an effort to construct systems which *embody* facets of human perception and cognition, such as chunking and auditory stream segregation (see Maxwell [2014] for a compelling example of this integrated approach). Although this approach may facilitate the evaluation of audience response, our aim is not necessarily to advocate for human-like systems (or to specify any particular setup of MuMe systems), but rather to examine perceptual, cognitive, and affective responses in humans as a means of evaluating the impact of the system's creative output. For an overview of the methods discussed in this section as well as the subsequent section on methods for internal evaluation, please refer to Table I.

Importantly, in addition to providing a means for researchers to test their creative system, the methods below *may be adapted to provide feedback within the system* (highlighted in Figure 1B/C). The implementation of feedback loops simply requires the system to incorporate or respond to the evaluated data collected. For example, real-time response data from the audience (on-line listener ratings, physiological measurements, motion capture data, etc.) may be used to update or influence the creative processing within a system. Criteria based on the response data could also be implemented for an internal reward system, for example, whether penalizing undesirable creations or dynamically weighting behaviors that produce a desired impact on audience states—there are precedents for this approach in CC [e.g., Saunders 2012].

4.1. Behavioral tests and the assessment of human creativity

Because we are addressing the evaluation of creativity, and creativity is a human construct, techniques for assessing creative processes and output should be applicable to both humans and machines. Given the long history of testing creativity in humans, we begin with an overview of behavioral tests of creativity. These techniques may be adapted by researchers to test the creativity of their MuMe system.

Behavioral tests of creative thinking are fairly common, as assessing creativity is relevant across a wide range of educational and professional contexts. Sets of behavioral tests (called batteries) are also commonly used for children's entrance exams, identifying gifted students, or placement into enrichment classes or speciality art schools. Most behavioral batteries test divergent thinking, convergent thinking, artistic ability, or rely on self-assessment. Divergent thinking is the ability to generate new ideas or solutions to some problem or task. Because divergent thinking can result in ideas that are not novel, metrics concerning originality, applicability, and value are often employed in conjunction with divergent thinking tasks to assess creativity. Conversely, convergent thinking tasks measure the individual's ability to discover a particular solution to a problem, and often require employing different strategies to find the solution. Self-assessment tasks prompt the individual to report his or her experience with creative pursuits, or ask questions about personality and creative inclinations. And lastly, artistic tests are usually domain-specific methods for testing proficiency or evaluating the creativity of an artistic object (such as a poem or musical score), often relying upon expert judgment. All of these assessment methods may be applied to the development or improvement of computationally creative systems.

The most common behavioral battery of creativity is the Torrance Tests of Creative Thinking, or TTCT [Torrance 1998] (developed in 1974 and re-normed in 1974, 1984, 1990, and 1998), and has been employed for testing infants up through adults. The TTCT primarily tests divergent thinking ability, and is comprised of verbal and figural problems. Measures of assessment for the original TTCT were based on traits from the model of Guilford [1959], including Fluency, Flexibility, Originality, and Elaboration. Fluency measures the number of generated ideas that are relevant to the problem (such as the number of figural images produced in a drawing task). Flexibility measures the similarity of responses, and whether solutions are based on multiple domains or ways of thinking. Originality refers to the number of responses that are statistically infrequent. And lastly, Elaboration measures the individual's ability to develop and add ideas to pre-existing ones [Torrance 1998; Kim 2006].

When re-normed, two new measures were added, and Flexibility was eliminated because it was found to be highly correlated with Fluency [Kim 2006]. This yielded the five measures of Fluency, Originality, Elaboration, Abstractness of Titles, and Resistance to Premature Closure. Abstractness of Titles refers to the degree of understanding and abstraction of thought beyond mere labeling of pictures. Resistance to Premature Closure measures the amount of "psychological openness" used when considering alternative ideas and processing the available information [Kim 2006]. In addition to these five measures, Torrance [1998] identified a set of thirteen subscales of "creative strengths", including, for example, richness of imagery, humor, and emotional expressiveness. The Torance tests are a prime example of testable measures of creativity that may be adapted as a metric for evaluating creative behavior (such as divergent processing and flexibility) and output (such as the originality of artefacts) from artificial systems.

Various other tests of divergent thinking, some of which were predecessors to the TTCT, are still often utilized as well. Most of these tests examine individuals' elaboration of possible solutions, measuring detail, quantity, novelty, and variety of ideas.

One approach by Wallach and Kogan [1965] has participants list as many items as possible that fall into a particular category. Another example is the Alternative Uses Task, developed by Guilford [1967], which asks participants to enumerate possible uses for a common object. There are also domain-specific divergent thinking tasks, which also fall under the category of artistic ability assessment. Many of these are also based on the work of Guilford and Torrance. The Measures of Musical Divergent Production [Gorder 1980], for example, tests musical participants on the number of improvised sounds produced (Fluency and Elaboration), shifts of musical content (Flexibility), Novelty, and "musical appeal". Because this assessment requires musical knowledge, scoring is performed by expert musicians, which is akin to the Consensual Assessment Technique discussed below. Systems designed to facilitate sound designers and composers in exploring complex sound synthesis parameter spaces explicitly aspire to notions expressed in domain-specific divergent thinking evaluation methods. In discussing the evaluation of *Genomic* [Stoll 2014], the author informally discusses the ability of the system to successfully explore the space of possibilities and generate high novelty populations, and move towards more rigorous methods of quantifying performance by considering the diversity of output sets in terms of similarity measures between generated sounds. For other examples of musical creativity tests, see the Measure of Musical Problem Solving [Vold 1986], and the Measure of Creative Thinking in Music II [Webster 1987]. If used within a MuMe performance setting, these musical evaluation criteria would lend credence to claims about the creativity, quality, and impact of the music generation system on listeners.

Convergent thinking tasks approach creativity in a slightly different manner. To successfully complete these tasks, creative thinking is required in terms of reassessing a problem, finding an insightful problem-solving perspective, or shifting strategies to hone in on the correct solution. An example of a convergent thinking task is the Remote Associates Test [Mednick 1962], in which participants must find one word that associates three given words. In terms of MuMe systems, a convergent task may be something like synchronizing the artificial system with the performer (by using various strategies based on rhythmic cues, pitch information, and shifts in harmony; see Cont [2011] for an overview of score following research), or finding converging means of eliciting a particular affective response in the listener, such as the rule-based system developed by Livingstone et al. [2010] designed to enable the real-time modification of musical structure and performance attributes for the purpose of inducing particular emotional states in the listener.

Self-assessment tests request the examinee to contemplate their own creativity, and are discussed below in Section 5.2 in the context of implementing techniques within the system for self-reflection. Lastly, as noted above, artistic tests often aim to assess the creativity of artefacts by recruiting expert opinions. The Consensual Assessment Technique may therefore be considered an implementation of this kind of test; we present an overview of this method below.

4.2. Consensual Assessment Technique

Csikszentmihalyi [1999] proposes that creativity entails a process that produces a new idea or artefact which is *recognized as creative by others*. Possibly the most widely applied method of external validation is called the Consensual Assessment Technique (CAT) [Amabile 1982, 1996a], in which a panel of experts from the relevant field judge the creativity of an artwork, theory, product, or performance. The method combines the critiques of the various judges, and is not based on any one particular theory of creativity [Amabile 1982]. The argument behind this approach is that the best reference for judging creativity comes from those with in-depth knowledge of the area. Because critics can have differing opinions, the assessments are pooled across judges.

That said, the technique has been shown to demonstrate reasonable reliability and reproducibility across sessions [Baer and McKool 2009]. One limitation of the CAT, however, is that the processes underlying the generation of the creative artefact are not considered.

It is worth drawing a comparison between that CAT and the (inaccurately) so-called 'Turing Test' methods of evaluation. There is a wide range of interpretations of this latter approach. For example, Cope [2005] has claimed success of the music generated by his system EMI in terms of its reception by a human audience (though not in a direct guessing game). However, there are some points to be made.

First, the Imitation Game, discussed in a thought experiment by Alan Turing [1950], requires a direct comparison by a player between a hidden human and a hidden computer in a task in which a very high level of performance would be expected of all healthy humans: general conversation. Musical composition is not, for whatever reason, such a task; this leaves most players ill-equipped to make informed judgments about the 'game'. Second, in Turing's Imitation Game, it is the aim of the hidden human to attempt to fool the player into choosing incorrectly; to our knowledge, this has not been attempted in this kind of study—and it is far from clear how an experimenter would go about doing so.

The CAT, on the other hand, requires experts for its comparison, addressing the first of these differences. Because it focuses only on one creative system, it does not address the attempt to deceive. Instead, and much more usefully in a scientific context, it requires qualitative feedback which can be used to enhance the creative system in question, as did Pearce and Wiggins [2007].

4.3. Extensions within computational creativity

Some of the behavioral methods and strategies listed in section 4.1 have been adapted for the evaluation of computational creativity. As discussed above, for example, a version of the CAT has successfully been used for assessing the creativity of generated music [Pearce and Wiggins 2007], an approach that is also useful for eliminating judges' bias against artefacts generated from artificial systems.

When expert evaluation is not possible, non-experts (as well as the researcher) may use a set of criteria for assessing the output of a potentially creative system. In the field of CC, Ritchie [2007] has proposed one such set of criteria, which are similar to the evaluation metrics used in the evaluation of human creativity. His empirical criteria are based on novelty (in terms of untypicality, innovation, or class membership) and quality (in terms of value ratings), and allow for variation in subjective judgment of creative artefacts as well as the variation in the criteria used to define creativity itself [Ritchie 2001]. Note that he too focuses on the generated artefact, and not on evaluating the internal processes underlying the creation of that artefact.

A framework for evaluating creativity that focuses both on generated artefacts and the creative behavior of a system is Colton's creative tripod [Colton 2008]. Colton emphasizes that knowledge of the creative process influences an observer's judgments of creativity, in value judgments of human creativity, but also and especially in the context of computational systems. He presents his evaluation methods with the knowledge that observers often demonstrate a bias against artificial systems: although exceptions exist, computationally-generated artefacts are often judged to be less creative than human generated ones, or the "real" creativity is attributed to the programmer, not the system [Moffat and Kelly 2006]. With this knowledge, the creative tripod is offered as a technique for describing and evaluating the behavior of creative systems, and is based on assessing skill (technical ability), appreciation (valued in the domain), and imagination (appropriate novelty that moves beyond pastiche) [Colton 2008]. Colton argues that if the system is considered skillful, appreciative, and imaginative, then the

software should be deemed creative. Colton and colleagues have also worked to begin formalizing computational creativity theory (CCT) by offering two descriptive models as a starting point: one focused on the act of creative generation, called FACE, and another which tests the impact that creative systems may have on the observer(s), called IDEA [Colton et al. 2011]. The FACE framework may be used to describe the processes underlying melodic/harmonic generation and rhythmic generation in MuMe models, and the IDEA approach may be used to incorporate the background knowledge and preferences of the audience.

4.4. Questionnaires, correlational studies, and rating scales

Surveys and questionnaires can be a valuable way for researchers to examine listeners' subjective responses to their system or their system's output. Assessment of this type may be based on the collection of factual information (e.g., collecting data about listeners' age or years of musical training may be incorporated so that the system reacts to and performs for specifically the present audience), or subjective responses from listeners. Open-ended questions may be employed, such as, "What aspects of this music strike you as novel and why?" Questions posed to listeners may be chosen to address specific research questions of interest, or they may be based upon the behavioral and psychometric tests above, or draw upon, for example, the set of key components of creative systems outlined in Jordanous [2012]. This qualitative, open-ended format enables the audience to freely express their opinion, but responses can be more difficult to pool across participants, especially in real-time. Surveys using discrete response criteria (e.g., asking the listener to select one of several options, as with multiple-choice questions) or rating scales may therefore be easier to implement for providing real-time feedback to the creative system. For a discussion of surveys and the evaluation of creative output in live performance settings, see Eigenfeldt et al. [2012].

Alternatively, responses to quantitative questions are well suited to be used to examine correlations between properties of the generated artefacts and their impact on listeners. Correlational studies use a statistical analysis to assess the relationship between two variables. The relationship is measured in terms of direction (positive or negative correlation) and strength (e.g., a correlation coefficient between 0 and 1). Although correlational studies do not provide evidence for causality, they can be a valuable indicator, and inspire empirical studies that can test causal relationships. A useful technique within computational creativity is to correlate properties of the creative system or the creative artefact with behavioral responses from observers. For example, listeners' ratings may be collected to assess cognitive and affective states in response to melodies varying in style or complexity. Several different types of response scales may be used for these quantitative assessments. A common tool is the Likert scale, a psychometric rating scale in which participants respond on a symmetric response range from disagreement to agreement. For example, a participant may be asked to respond on a scale from 'Strongly Agree' to 'Strongly Disagree' to capture how strongly they believe a particular statement is valid. Other rating scales are continuous (as used by Egermann et al. [2013]), or partition the scale numerically, asking, for example, "How creative [enjoyable/interesting/novel] was this melody on a scale from 1 to 7?" Scales of this sort have been utilized to prompt listeners about preference, expectation, perceived complexity, aesthetics, functionality, novelty, and more. Although these are subjective tests, when administered correctly, ratings scales can provide very robust and consistent measures of participants' judgments (e.g., Preston and Colman [2000], although Yannakakis and Martínez [2015] give a precautionary account).

For researchers concerned with the individual subjectivity of observers' ratings, group-level approaches such as the Consensual Assessment Technique [Amabile 1982], described in Section 4.2, may be of use. By measuring the degree to which a group of

assessments agree, the issue of individual subjectivity is resolved. The CAT in particular also utilizes measures of reliability (consistency in agreement among raters) and validity to ensure that the findings are robust and replicable. For these reasons, MuMe researchers may find averaged responses (across a group of observers, especially experts) very helpful for systematic and reliable evaluation of their systems.

4.5. Measuring movement and physiological response

Physiological measurements can be used to capture physical manifestations of psychological and emotional states, thereby making these methods very relevant for MuMe performance systems. Common measures are heart rate, or Blood Volume Pulse (BVP), respiration, and Galvanic skin response. Skin conductance is an electrodermal response indicative of a person's physiological and psychological state of arousal (a term used to describe overall activity or energy). Another useful method is electromyography (EMG), which measures small movements in the muscles associated with smiling and frowning. These techniques have been used to assess perceived tension, stimulus intensity, or evoked emotion in a range of domains, and have also recently been applied to measuring audience physiological and affective response during music listening [Egermann et al. 2013]. MuMe systems could also use physiological measures to capture listener levels of engagement, and to model an audience state for the purpose of providing feedback to the computational system. Enabling creative systems to dynamically learn, model and react to the audience can eventually lead to systems with self-reflection, which of course will impact the way in which the computational system performs.

4.5.1. Motion capture. Motion capture is another approach to measuring real-time indicators of emotion, arousal, and embodied cognitive states. In this technique, motion sensors are placed on the participant's body to record movement in real-time while performing a task. This technique has been used in studies of music performance, for example, to correlate performers' movements (sometimes during specific sections of music) to perceived emotion in the audience [Livingstone et al. 2009; Friberg 2004]. In the context of a MuMe performance setting, researchers could use this method to capture performer or audience movements (again, indicating affective states), and examine whether movement differed between different types or sections of generated music.

4.5.2. Eye-tracking. Eye-tracking is another technique that indirectly measures online, real-time aspects of attention, perception, and cognition. Eye-tracking may be used as a measure of attention, information-processing, and decision-making Hayhoe and Ballard [2005]; Duchowski [2002], as well as exploratory behavior (by measuring saccades around an image or visual scene) Conati and Merten [2007]. For MuMe systems that collaborate with a human performer, this method may be used to elucidate the performer's perceptual and cognitive states, which may also be built into the system as feedback. For example, if a pianist is co-creating with a computational system, measuring his eye gaze may provide information about the notes that the pianist is considering playing during real-time performance, or alternatively, the moments in which the music does not line up with the performer's expectations. A benefit of this method is that no overt response is needed; the participant may go about their task at hand in a relatively undisturbed manner while data are collected. Eye-tracking may also be useful for multi-modal MuMu performances, to assess visual attention and engagement in the audience.

4.5.3. Reaction time. Rather than ask participants to make direct judgments or ratings, reaction time (RT) may be used as an indirect measure of perceptual or cognitive

processing. RT is the time between the presentation of a stimulus and a subsequent behavioral response, such as a button press. This method may be used for a variety of tasks; for example, RT data may be used to assess the relative expectedness of stimuli that vary in predictability or group membership, with the hypothesis that slightly altered stimuli will result in longer reaction times than grossly deviant stimuli [see for example Wang et al. 1994]. In the setting of a MuMe performance, a listener may be asked to respond when a rhythmic or melodic pattern changes, to inform the researcher about the listener state. Alternatively, if a co-creator or musician is working with the MuMe system, incorporating RT may yield a more robust model tailored to the co-creator's reactions.

4.6. Measuring neural responses using electroencephalography

For those MuMe researchers interested in incorporating the participant or listener's neural activity into their system or performance, electroencephalography (EEG) may be the most viable technique. EEG measures the electrical activity in the brain, as measured on the scalp. Populations of neurons, when firing together, emit enough electrical activity to be measured (albeit amplified) through the layers of meninges and cranium. It should be noted that EEG measures cortical activity; electrical signals from inner brain structures (such as the basal ganglia) dissipate too much by the time they reach the skull. Changes in electrical activity are thought to reflect underlying post-synaptic neural processing, with particular kinds of EEG oscillatory activity, in different parts of the brain, indicative of particular types of perceptual or cognitive processing. EEG methods are often not used for localization of function, as the spreading signal is broadened and muffled by the time it reaches the electrodes on the scalp. Rather, EEG has excellent temporal resolution, with measurements (on as many as 256 electrodes placed around the scalp) recorded as often as every 1 or 2 milliseconds. Because of the fine-grained temporal measurement, and the modern availability of affordable and transportable EEG caps (such as the Emotiv system), electroencephalography is well-suited for music research, and has been used with performers during live music generation performances and in the context of brain-computer music interfacing systems [Grierson and Kiefer 2011, 2014; Miranda 2006; Rosenboom 1990].

One of the most common EEG techniques is event-related potential (ERP) analysis, which measures electrical activity (in terms of amplitude in μV and latency in ms) immediately following an event (i.e., an experimental stimulus). There are characteristic waves in the ERP response: First the N1 is a large negative-going evoked potential, which peaks in amplitude around 80-120 ms post-stimulus onset. This activation is usually distributed in frontal-central areas. This wave is followed by a positive-going evoked potential (the P2), which peaks around 150-270 ms post-stimulus onset. Researchers make a distinction between the auditory evoked response and visual responses, for example, which display small variations in amplitude and latency. The N1 is a preattentive response that is sensitive to the unexpectedness of an event [Lange 2009], as well as physical properties of the stimulus (in the case of audition, the loudness or frequency of the stimulus, for example). In addition to stimulus predictability, the amplitude of the N1 can also reflect focused attention [Luck et al. 1990; Alho et al. 1997].

In terms of music generation and creativity research, ERP components can be used to assess expectation mechanisms and semantic processing in the audience or co-performer. As discussed above, the amplitude and latency of the N1 component is modulated by unexpected events occurring within a predictable context. Because novelty is an important aspect of evaluating creativity, a neural response to unexpected stimuli is a useful tool for assessing perceived surprise. Evidence also exists that N4 is sensitive to predictability: the amplitude of these components may increase during

rule-learning, as when participants are implicitly learning the transitional probabilities between tones within a "tritone word" (a set of three tones, analogous to a word with three phonemes) [Abla et al. 2008]. The N4, then, may be used as a proxy for statistical learning or segmentation. Because segmentation ability can reflect the amount of implicit knowledge acquired, as well as the training/experience of the individual, this measure can be used for comparisons of computational segmentation and statistical mechanisms with neural signatures of learning and surprise in humans.

Researchers also commonly examine temporal and spatial activation over broad swaths of the cortex. In time-frequency analyses, global oscillatory activity is assessed in terms of different bands of activity. The theta band (4–7 Hz) is seen during drowsiness and states of low physiological arousal, but has also been connected to suppression and inhibition. Alpha activity (8–12 Hz) has been widely studied, with some controversy of findings. It is commonly associated with calm states – posterior alpha waves (distributed over the occipital lobe) while the individual's eyes are closed is often indicative of relaxation or drowsiness. Like theta activity, some research argues that more alpha power reflects passive or active inhibition of non-relevant information while performing a task [Worden et al. 2000]. Finally, beta band activity (13–30 Hz) typically reflects active, alert states. In addition to measuring these oscillatory bands to use as real-time feedback for MuMe systems, oscillatory activity has also been used to guide creative interaction between the system and the co-creator [Miranda 2006].

An approach within the neuroscience of creativity has been to examine oscillatory band activity during creative compared with non-creative tasks. One hypothesis is that alpha activity supports divergent thinking [Martindale et al. 1990], and therefore more alpha and less beta band power should be present while individuals perform a creative task. Although this hypothesis has found reasonable support [Jaušovec 2000; Martindale et al. 1984], it should be noted that some studies have not seen this elevation in alpha power for creative versus non-creative activities (see Dietrich and Kanso [2010] for a review). In sum, oscillatory activity may be used with musicians or co-creators during real-time performance [Rosenboom 1990; Miranda 2006], or to capture the audience's affective state [Egermann et al. 2013] and incorporate it into the musical creation.

4.7. Benefits and limitations of particular methods

Choice of external methods may rely on several factors, including the goals of the researcher for the creative system, the architecture of the system, and even budgetary and time constraints. It is therefore useful to consider the pros and cons of the various methods mentioned above. The alternatives for behavioral assessment of creativity (e.g., the Torrance tests) may be easily adapted to ask listeners directed questions about their perception and emotional responses to the creative system. Indeed, questionnaire data can be an attractive resource for MuMe researchers because these data are relatively easy to acquire and analyze, and no expensive, specialized equipment is needed for data collection. As discussed above, correlating properties of creative output with listeners' responses can both help researchers understand what is and is not working for their system, and be used as audience feedback to the system. A potential drawback, however, to collecting behavioral responses, especially from non-experts, is that the act of asking observers to rate their views can actually alter the observers' views [Schwarz 1999]. Because response bias may be reduced in experts, the researcher may seek to obtain experts' assessments of the creative artefacts using the CAT, although this can require significant time and funding. Methods drawn from computational creativity, such as the creative tripod [Colton 2008, ; see §4.3], may also be particularly helpful, as these methods have been specifically devised for testing computationally creative artefacts.

An alternative approach to collecting behavioral responses is to assess creativity through indirect measures of perception and cognition, such as eye-tracking and reaction time measurements. These methods can skirt the issue of bias, and are also useful in cases where expert evaluation is not feasible. The drawback, in terms of ease of use, is that special equipment, as well as specialized software for data analysis, is necessary to implement these physiological methods.

5. INTERNAL EVALUATION: MODELING HUMAN CREATIVE BEHAVIOR AND AUDIENCE PERCEPTUAL, COGNITIVE, AND AFFECTIVE STATES FOR SELF-REFLECTION

Methods for modeling human perception, cognition, and affective states, together with their associated methods of evaluation, can inform the development of more sophisticated creative systems endowed with self-evaluative capacities. Self-evaluation, at some level, may be regarded as a necessary component of creative acts. Human creators make judgments regarding artefacts, creative processes or affective responses, both from their own knowledge and perspective, but also from an understanding of how their practice or works may be received. Such internal evaluative mechanisms may simply provide direct feedback for the creative process, leading to the refinement of artefacts. On a more sophisticated level, self-evaluation can form the basis of reflection, whereby the the entire creative process itself becomes the subject of consideration. Self-reflective agents may be capable of exhibiting creative behavior that corresponds with what Boden describes as transformational creativity, which in terms of the CSF (section 2) is redefined as "exploratory creativity at the meta-level" [Wiggins 2006a, p. 6].

Reflection may be based on the methods already presented above, e.g., a system may employ a model of audience affective responses which is based on the collection of EEG data, but note that the data itself is not sufficient for reflection; the system must have a means of reasoning about the data. Alternatively, the system may employ, in addition to its creative processing, a model of some aspect of the audience (and their background/context, such as the musical experience of the target audience) to reflect upon the impact of its creative behaviors. A complete survey of the subject of reflection within CC is beyond the present scope, however, we discuss below some prominent techniques for modeling human perception and cognition, which may be used to enable creative systems to incorporate knowledge of their audience into their reasoning and meta-reasoning processes. We begin by discussing how the methods presented in Section 4 may be used as the basis for self-reflection within a creative system.

5.1. Self-assessment based on behavioral tests

Section 4.1 above included a discussion of how behavioral tests, such as the Torrance Tests of Creative Thinking, may be adapted to allow researchers to externally test the creativity of their system. Divergent and convergent thinking tests may also be adapted as *internal* tests for self-reflection. Consider a system that employs an internal model of measures such as Originality and Elaboration, or uses subscales of creative strengths [see Torrance 1998] such as emotional expressiveness. If the system has a set of self-evaluation metrics such as these, and a means of assessing itself via these criteria, the system is then able to situate its own behaviors and output in the context of what it aims to achieve artistically, and how successful it may have been.

Self-assessment tests offer another clear way to draw from human behavioral tests for the purpose of self-reflection. These tests implore the examinee to reflect on their own creativity, or report autobiographical information, such as the participant's involvement in artistic pursuits (e.g., experience taking a visual arts class or writing a short story). Self-assessment may be based on an existing conceptual framework of the creative process, such as that suggested by Wallas [1926], which includes the

four stages of preparation, incubation, illumination, and verification. Many formal self-assessment tests are also based on the work of Torrance, such as the Khatena-Torrance Creative Perception Inventory [Khatena and Torrance 1976]. Some work in the field of CC aims to prompt creative systems to examine and critique their own output [Colton et al. 2014], but this has not yet become standard practice in MuMe.

5.2. Reflective models based on external tests of audience perception, cognition, and affective states

Data from perceptual and physiological external evaluation methods may be used as the foundation for internal reflection if the creative system reasons about the data. For example, the system may form expectations about how its artefacts will be perceived by the audience, tailor its creative processing to produce a desired effect (e.g., an affective response in the listener), and, given these expectations and goals, subsequently reflect upon whether the desired effect was produced in the listener and why. If a system strives to invoke a particular type of affective response in the listener, it should possess some sort of internal model of human affective responses (based, for example, on the two-dimensional model of valence and arousal). Then, once feedback from humans is collected (e.g., questionnaire ratings are recorded) and analyzed with regard to the internal model of affective response (e.g., average scores on ratings scales indicate particular affective states in the audience), the system may reflect on whether its target was met. Regardless of the type of external feedback received, if the system is given a means of predicting how listeners are likely to respond, it may compare listeners' actual responses with its predictions.

In contrast to post-performance reflection, real-time measures of physiological responses may be used for continuous, on-line self-assessment and reflection. For example, motion capture data that is fed back to the system may provide valuable information about whether the generated music is resulting in the particular movement responses (e.g., tapping along to the beat or dancing to a particular rhythm) that the system hopes to elicit in the audience [e.g., Solberg 2014]. If systematic changes in the audience are not discovered, the system may want to alter its creative process. Along the same lines, eye-tracking measurements, reaction time data, heart rate data, and continuous EEG may be used by the creative agent to assess whether the desired goals of the system have been achieved, and whether (and how) the creative processes should be updated to support future desired outcomes.

5.3. Measures of prediction and expectation for modeling perception and affective response

Temporal processing of sequences, such as music and language, relies heavily upon expectation and prediction mechanisms. Our brain picks up statistical relationships and patterns in our environment from which we learn and make predictions about the world [Saffran et al. 1996; Creel et al. 2004; Lew-Williams and Saffran 2012]. Because expectation is crucial for learning, and influences what enters perceptual awareness as well as what is encoded in memory, it may therefore also be considered integral to human creativity [Wiggins 2012b]. While listening to speech, one does not passively "hear" words go by; rather, the listener makes implicit predictions about what will happen next. For example, in the sentence, "I saw the bird flap its…", the listener will have a strong expectation for "wings". If an unexpected word occurs instead, such as "banana", a surprise response to the semantic incongruity is elicited in the brain [e.g., Kutas and Hillyard 1984]. Through correct and incorrect expectations (that is, expectations that are either validated or violated by the ongoing stream of information), we learn the rules and structure of the signal or stimulus. We are then able to form increasingly robust and accurate prediction mechanisms, and the learning process continues dynamically in this cycle of prediction, feedback, and updating of an

internal predictive model [Clark 2013; Garrido et al. 2007]. Relating this perspective on human cognition to computational creativity, researchers may consider using a similar type of prediction-action feedback loop in their MuMe systems [c.f., Brown and Gifford 2013], as this mechanism enables humans to learn a statistical framework and network of associations (analogies, scripts, schemas, etc.) that guide perception, cognition, and creation.

Correct prediction is a valuable evolutionary trait, and because we cannot always rely on direct experience to form predictions (for example, when encountering an unfamiliar member of the large cat family), theorists posit that humans have developed affective responses to the validation or denial of predictions, even in domains such as music in which there are no real threats [Huron and Margulis 2010; Huron 2006]. Through this emotional response to prediction, we develop preferences about what we encounter in the world. The incorporation of models of listener expectation into generative procedures could make a powerful impact to the expressivity and creativity of MuMe systems. Livingstone et al. [2010] identify the modeling of expectation and tension as a future step in the development of their Computational Music Emotion Rule System. Gaining the ability to predict likely listener responses to new material in terms of expectation enables systems to explicitly model higher-level compositional intentions concerning the elicitation of particular affective responses–for example, Huron [2006] and Egermann et al. [2013] provide potential inspiration.

In relation to phenomena such as novelty seeking, creativity, and aesthetics, hedonic preferences may often be described in terms of an inverted-U relationship as described by Wundt and Berlyne [Berlyne 1970]. The hedonic function stipulates that as stimulus intensity or complexity increases, preference or pleasure increases until the top of the Wundt curve, after which increasing complexity results in diminishing preference. Essentially, theory posits that stimuli that are very predictable are perceived as boring, and stimuli that are extremely complex or intense are viewed as inaccessible or over-stimulating. Therefore, the center of the curve yields a “sweet spot” of optimal complexity, which is often both a function of the complexity/intensity of the stimuli, as well as the background and experience of the observer. This theory has been applied in a range of domains, including music perception and creativity research [e.g., Steck and Machotka 1975; Martindale et al. 1990; Saunders and Gero 2001; North and Hargreaves 2008].

The role of expectation is not new to the field of creativity, as demonstrated by the work of Wiggins and colleagues and others [Wiggins et al. 2009; Pearce and Wiggins 2007; Grace and Maher 2014; Maher 2010]. Pearce and Wiggins have used measures of unexpectedness in implementations of computational systems and in behavioral tests of music perception [Pearce and Wiggins 2006; Pearce et al. 2010; Agres et al. 2013]. Maher and colleagues speak of expectedness and surprise as an evaluation metric—like many other approaches, novelty and value are identified as mandatory criteria for evaluating creativity, but in addition to these, Maher adds unexpectedness. She argues that whereas novelty is a metric for identifying how different a particular artefact is from other examples in its class, unexpectedness is based on a set of temporal expectations; it quantifies not how different the artefact is, but how much it deviates from expectation [Maher 2010].

Researchers looking for explicit techniques to build expectation into their systems may want to turn to information theory, which provides tools for implementing measures of expectation and prediction as discussed above. Information theory measures the amount of information contained in a transmitted signal, and has been widely applied in fields as diverse as astronomy and linguistics. Measures such as entropy (amount of uncertainty) and information content (IC: unexpectedness; the probability of an event given context) have been shown to reflect measures of perceptual and

cognitive processing. For example, in music perception, Pearce and colleagues have shown that IC captures the unexpectedness of musical events [Pearce and Wiggins 2006; Pearce et al. 2010; Agres et al. 2013]. By calculating the IC of discrete events, one can make predictions about which events should be the least or most expected in the sequence [Pearce et al. 2010; Pearce and Wiggins 2012]. For behavioral validation, participants can rate the expectedness of the events, as this has been shown to be a robust measure for capturing participants' psychological expectation and surprise. Rather than solely examine point-wise measures of information, one can also test the effect of information properties of the entire sequence [Agres et al. 2013]. The information theoretic properties of whole sequences have been shown to have a dynamic effect on perception and memory over time, with complex stimuli often having an increasing effect of poor recognition memory performance. As discussed above, expectation mechanisms play a prominent role in listeners' perception and experience of music. By using information theoretic measures and behavioral testing, MuMe researchers may quantify listeners' expectations and then apply this knowledge in generative systems to influence listeners' perception of melody and rhythm, create a desired affective response, and model the listener's mental state.

No summary of information theoretic approaches to creative cognition would be complete without mention of the work of Jürgen Schmidhuber—in particular, his Formal Theory of Creativity, Fun, and Intrinsic Motivation [Schmidhuber 2010]. This relates to the important question of *why* humans like to be creative, even when doing so does not benefit them in any biological way. The account is based on an evolutionary advantage afforded by abilities to identify patterns in the world (and thus be information-efficient), and to compress the overwhelming amount of data which assaults an organism in the process of normal perception. Hedonic reward is hypothesized to arise from successful compression, providing a notion of 'beauty', though the definition of beauty as compressibility is, in our option, overly simple. Biederman and Vessel [2006] supply an endocrine mechanism via which Schmidhuber's theory might be implemented.

To summarize the relevance of expectation mechanisms on computational creativity: prediction helps the brain learn and encode information about a domain, and (violation of) expectation is related to aesthetics and affect. This may be modeled, at least in part, using information theory. Creativity often involves the discovery of novel solutions to a problem or task; it is the act of exploring (finding new regions or pathways) or extending a learned space of mental representations. The network of mental representations or conceptual spaces may be thought of as a complex prior distribution, with statistically-defined co-occurrence and correlational features. In this framework, exploratory creativity involves generating new ideas based on learned probability distributions or conceptual representations, and preference for generated creative artefacts (value judgments) generally falls within a preferred range of familiar and novel, or predictable and complex.

5.4. Conceptual representation

Given the importance of conceptual blending, associative processing, and the novel combination of existing ideas or examples within the field of computational creativity, it is useful to consider frameworks which specifically address the structure and properties of concepts themselves. While several approaches address this point [Barsalou 2008; Barsalou et al. 2003; Bareiss 2014], we focus on Peter Gärdenfors' conceptual spaces theory [Gärdenfors 2000], because it provides a quasi-formal framework with stated mathematical principles to underpin the cognitive modeling.

Gärdenfors [2000] presents the theory of conceptual space as a representational tool for approaching problems concerning the modeling and understanding of cognitive processes. As a representation, it is situated at a particular level of abstraction.

He argues that conceptual structures should be represented using geometry on what he terms the *conceptual level*. This level of representation is situated between the *symbolic level*, which includes, for example, formal grammar, and the *sub-conceptual level* of high-dimensional representations such as neural networks. The theory states that concepts, which are entirely mental entities, can be represented within sets of *quality dimensions* with defined geometrical, topological or ordinal properties. These dimensions correspond to ways of comparing objects, namely, how similar (proximate) or distant stimuli are within a space. Thus, similarity between concepts is implicitly represented in terms of the distance between points or regions in a multidimensional space, in a manner comparable to the spatial view of similarity proposed by Shepard [Shepard 1962a,b].

For example, a conceptual space for color is defined by the three dimensions of hue, saturation, and brightness, all of which can be described in terms of geometrical shapes. Certain dimensions are integral; this signifies that giving value to one dimension (e.g., pitch) necessitates a value in another dimension (loudness)—if a pitch exists, it must have a certain loudness. Other dimensions are separable, such as instrumentation and rhythm. Gärdenfors defines a domain as a collection of integral dimensions that are separable from every other dimension. This yields a slightly more specific definition of a conceptual space as a set of quality dimensions that are divided into domains. Additional formalizations are present to define the geometrical characteristics of domains and regions, such as convexity of natural properties. For a more in-depth account, see Gärdenfors [2000].

One can apply conceptual spaces theory to both the processes underlying creative generation and to modeling and incorporating the state of the audience or performer whilst creating. For example, given an appropriate conceptual space, it is possible to develop hypotheses, specified in terms of geometry, about how novel musical ideas are likely to be perceived by the listener. This is analogous to the estimation of likelihood given a stochastic model. However, in addition, the geometry affords a particularly intuitive means of interpretation and reasoning, accessible to both humans and machines. Understanding can be modeled in terms of proximity, spatial regions, trajectories and transformations, and other properties inherent in the geometrical structures [c.f. CSF, Wiggins 2006b]. In practical musical terms, potential undertakings could include constructing conceptual space representations of high-level concepts such as genre or mood, which could be used as the basis for blending between different styles, or as a component of a system that attempts to produce specific affective responses. Other possibilities might include utilizing perceptually validated feature spaces of melodic or rhythmic structure, which afford the grounding of concepts like “singalongability” or “danceability” in musical structure. Such spaces could be used by generative processes within the creative system to guide the production of new artefacts according to predefined criteria, or in a more autonomous and strongly creative system, the formation of the system’s own higher-level goals and intentions. Conceptual spaces can potentially allow MuMe systems, and their human collaborators, to explore and evaluate novel artefacts in terms of meaningful shared conceptualizations. These conceptual structures can support the modification of generative processes, either by the system designer or the system itself, to enable the exploration of previously uncharted conceptual regions, whilst importantly retaining the ability to relate new discoveries to already known forms of musical expression.

Significant challenges to the practical application of conceptual spaces theory are, firstly, the construction of such geometrically-meaningful spaces, and secondly, establishing their perceptual validity. A detailed discussion of this topic beyond the scope of the present review, and is itself a major area of research in the fields of cognitive science, psychology, and machine learning. Furthermore, as a theory of representation,

conceptual spaces theory can be combined with established modeling techniques, such as statistical models [for example, Pearce 2005] or deep neural networks [for example, Bickerman et al. 2010], to enhance the capacity of the system to begin to approach issues of meaning, in the perceptually grounded sense, of the data over which they operate.

One testable hypothesis that could be formed regarding creativity and conceptual spaces is that novel ideas/examples that lie within a conceptual space are less cognitively demanding to process (although also possibly perceived as less creative) than novel examples that change the geometry of the space (as in transformational creativity). Incorporating a conceptual spaces model of the audience or co-creator may supply the MuMe system with a means of self-reflection and theory of mind that may enhance the appropriateness and value of the generated music. This may be considered to be an anthropocentric approach to creativity, but again, because music (and musical creativity) is a human construct, the ways in which humans conceptualize of music can lend important insights for AI and machine creativity.

Some systems already incorporate methods inspired by conceptual spaces for creative generation or to serve as models of listeners' perception [Bååth et al. 2014]. One system, presented as a cognitive architecture, suggests the use of a formal implementation of Gärdenfors' conceptual spaces to draw analogies between different domains, namely musical pitch perception and visual scene perception [Chella 2015]. Forth [2012] presents a number of geometrical representations of musical structure designed to enable the processing of melodic and metrical-rhythmic structure in terms of perceptual similarity. Another approach facilitates exploratory behavior of conceptual spaces in the context of creative story generation [Veale et al. 2010]. Because an artificial system may generate myriad possible examples, many will inevitably be of poor quality. Therefore, a key component of the system above is to reject candidate stories by confining the system's exploratory behavior to more highly-valued regions of spaces; in other words, exploration of conceptual space is constrained based on quality dimensions, such as emotion, interest, and tension [León and Gervás 2010]. This technique may be applied to MuMe systems as well for exploring valuable new melodic, harmonic, and rhythmic material.

6. SUMMARY AND CONCLUSION

This paper provides an overview of theoretical and empirical methods that may be employed or adapted for the study of computational creativity, specifically for the MuMe community. At the broadest level we have addressed the issue of evaluation from two perspectives. Firstly, from a scientific standpoint, we have discussed a range of methods offering means of objectively assessing creativity. These methods can be employed by researchers to answer questions about the behavior or artefacts produced by artificial creative systems. We term this form of evaluation *external*, because the source of judgments or measurements is from outside of the system itself. The second perspective from which we have considered evaluation is from the standpoint of a creative system itself, and its *internal* capacity for self-evaluation.

In parallel with discussing methods for external and internal evaluation, we have highlighted connections to theories of creativity and computational creativity, and the broader subject of how making judgments about creative artefacts and the processes that produce them relates fundamentally to questions of system design. We provide a simple taxonomy (as depicted in Figure 1) which distinguishes between different system architectures based on the degree to which feedback and/or reflection are integrated with the generative capabilities of the system. Upon marrying the theoretical argument for the inclusion of both external and internal evaluation methods with the goals of the creative system, we frame our discussions within the context of four main

areas of interest to MuMe, including melodic and harmonic generation, rhythmic generation, creating an affective response in the listener, and incorporating performer and audience feedback.

Above all, this paper advocates a scientific approach to the study of creativity and the development of MuMe systems. Precise means of evaluation, in conjunction with unambiguously stated hypotheses, are a fundamental building block for advancement in any scientific or scientifically-aligned discipline. We have attempted to demonstrate the benefits of placing evaluation center-stage in that it provides benefits not only for individual researchers, but also the community as a whole. Furthermore, the subject of evaluation in creativity more generally offers insights for the development of more advanced creative systems in which a capacity for evaluation becomes a fundamental component.

ACKNOWLEDGMENTS

All three authors are funded by the project Lrn2Cre8, which acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 610859.

REFERENCES

D. Abla, K. Katahira, and K. Okanoya. 2008. On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience* 20, 6 (2008), 952–964.

K. Agres, S. Abdallah, and M. Pearce. 2013. An Information-Theoretic Account of Musical Expectation and Memory. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Eds.). Cognitive Science Society, Austin, Texas, 127–132.

K. Alho, C. Escera, R. Díaz, E. Yago, and J. M. Serra. 1997. Effects of involuntary auditory attention on visual task performance and brain activity. *Neuroreport* 8, 15 (1997), 3233–3237.

T. Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology* 43 (1982), 997–1013.

T. Amabile. 1996a. *Creativity in context*. Westview press.

T. M. Amabile. 1996b. *Creativity in Context*. Westview Press, Boulder, Colorado.

C. Anderson, A. Eigenfeldt, and P. Pasquier. 2013. The Generative Electronic Dance Music Algorithmic System (GEDMAS). In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

C. Ariza. 2009. The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems. *Computer Music Journal* 33, 2 (20 May 2009), 48–70.

G. Assayag. 1998. Computer Assisted Composition today. In *Proceedings of the First Symposium on Music and Computers*. Corfu, Greece.

G. Assayag, C. Rueda, M. Laurson, C. Agon, and O. Delerue. 1999. Computer-Assisted Composition at IRCAM: From PatchWork to OpenMusic. *Computer Music Journal* 23, 3 (sep 1999), 59–72.

R. Bååth, E. Lagerstedt, and P. Gärdenfors. 2014. A prototype-based resonance model of rhythm categorization. *i-Perception* 5, 6 (2014), 548–558.

J. Baer and S. McKool. 2009. Assessing creativity using the consensual assessment. *Handbook of assessment technologies, methods and applications in higher education* (2009).

R. Bareiss. 2014. *Exemplar-based knowledge acquisition: A unified approach to concept representation, classification, and learning*. Vol. 2. Academic Press.

L. W. Barsalou. 2008. Situating concepts. *Cambridge handbook of situated cognition, ed. P. Robbins & M. Aydede* (2008), 236–63.

L. W. Barsalou, W. K. Simmons, A. K. Barbey, and C. D. Wilson. 2003. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences* 7, 2 (2003), 84–91.

B. Bel and J. Kippen. 1992. Bol Processor Grammars. In *Understanding Music with AI – Perspectives on Music Cognition*, O. Laske, M. Balaban, and K. Ebcioğlu (Eds.). MIT Press, Cambridge, MA, 366–401.

D. E. Berlyne. 1970. Novelty, complexity, and hedonic value. *Perception & Psychophysics* 8, 5 (1970), 279–286.

J. Bharucha, M. Curtis, and K. Paroo. 2009. Musical communication as alignment of non- propositional brain states. In *Language and Music as Cognitive Systems*, R. M. C. I. Rebuschat, P. and J. Hawkins (Eds.). Oxford University Press.

G. Bickerman, S. Bosley, P. Swire, and R. Keller. 2010. Learning to create jazz melodies using deep belief nets. In *First International Conference on Computational Creativity*.

I. Biederman and E. A. Vessel. 2006. Perceptual Pleasure and the Brain. *American Scientist* 94 (May–June 2006), 247–53.

J. A. Biles. 1994. GenJam: A Genetic Algorithm for Generating Jazz solos. In *Proceedings of the 1994 International Computer Music Conference*. ICMA, San Francisco, CA.

M. A. Boden. 1998. Creativity and artificial intelligence. *Artificial Intelligence Journal* 103 (1998), 347–356.

M. A. Boden. 2004. *The creative mind: Myths and mechanisms*. Psychology Press.

T. L. Bolton. 1894. Rhythm. *The American Journal of Psychology* 6, 2 (1894), 145–238.

A. Brown and T. Gifford. 2013. Prediction and Proactivity in Real-Time Interactive Music Systems. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

A. R. Brown, T. Gifford, and R. Davidson. 2015. Techniques for Generative Melodies Inspired by Music Cognition. *Computer Music Journal* 39, 1 (2015), 11–26. Volume 39, Number 1, Spring 2015.

A. Burnett, E. Khor, P. Pasquier, and A. Eigenfeldt. 2012. Validation of harmonic progression generator using classical music. In *Proceedings of the Third International Conference on Computational Creativity (ICCC 2012)*. Dublin, Ireland, 126–133.

A. Chella. 2015. A Cognitive Architecture for Music Perception Exploiting Conceptual Spaces. In *Applications of Conceptual Spaces*, F. Zenker and P. Grdenfors (Eds.). Synthese Library, Vol. 359. Springer International Publishing, 187–203.

C.-H. Chuan and E. Chew. 2011. Generating and Evaluating Musical Harmonizations That Emulate Style. *Computer Music Journal* 35, 4 (2011), 64–82. Volume 35, Number 4, Winter 2011.

A. Clark. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36, 03 (2013), 181–204.

N. Collins. 2008. Infno: Generating Synth Pop and Electronic Dance Music On Demand. In *Proceedings of the International Computer Music Conference*. International Computer Music Association, San Francisco, CA, US.

S. Colton. 2008. Creativity Versus the Perception of Creativity in Computational Systems.. In *AAAI Spring Symposium: Creative Intelligent Systems*. 14–20.

S. Colton, A. Pease, and J. Charnley. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the Second International Conference on Computational Creativity*. 90–95.

S. Colton, A. Pease, J. Corneli, M. Cook, and T. Llano. 2014. Assessing progress in building autonomously creative systems. In *Proceedings of the Fifth International Conference on Computational Creativity*.

S. Colton and G. A. Wiggins. 2012. Computational Creativity: The Final Frontier?.

In *Proceedings of 20th European Conference on Artificial Intelligence (ECAI) (Frontiers in Artificial Intelligence and Applications)*, L. D. Raedt, C. Bessire, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, and P. J. F. Lucas (Eds.), Vol. 242. IOS Press, Montpellier, FR, 21–26.

C. Conati and C. Merten. 2007. Eye-tracking for user modeling in exploratory learning environments: An empirical evaluation. *Knowledge-Based Systems* 20, 6 (2007), 557–574.

A. Cont. 2011. On the creative use of score following and its impact on research. In *SMC 2011 : 8th Sound and Music Computing conference*. Padova, Italy.

D. Cope. 1987. Experiments in Musical Intelligence. In *Proceedings of the International Computer Music Conference*. San Francisco.

D. Cope. 1991. *Computers and Musical Style*. Oxford University Press.

D. Cope. 2005. *Computer Models of Musical Creativity*. MIT Press, Cambridge, MA.

H. Cowell. 1930. *New Musical Resources*. Cambridge University Press.

S. C. Creel, E. L. Newport, and R. N. Aslin. 2004. Distant melodies: statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, memory, and cognition* 30, 5 (2004), 1119.

M. Csikszentmihalyi. 1999. 16 Implications of a Systems Perspective for the Study of Creativity. In *Handbook of creativity*. Cambridge University Press, 313–335.

T. DeNora. 2000. *Music in everyday life*. Cambridge University Press.

A. Dietrich and R. Kanso. 2010. A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological bulletin* 136, 5 (2010), 822.

A. T. Duchowski. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers* 34, 4 (2002), 455–470.

K. Ebcioğlu. 1988. An Expert System for Harmonizing Four-Part Chorales. *Computer Music Journal* 12, 3 (1988), 43–51.

K. Ebcioğlu. 1990. An Expert System for Harmonizing Chorales in the Style of J. S. Bach. *Journal of Logic Programming* 8 (1990).

H. Egermann, M. T. Pearce, G. A. Wiggins, and S. McAdams. 2013. Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience* 13, 3 (2013), 533–553.

A. Eigenfeldt. 2007. The creation of evolutionary rhythms within a multi-agent networked drum ensemble. In *Proceedings of the International Computer Music Conference (ICMC 2007)*. Copenhagen, Denmark, 3–6.

A. Eigenfeldt, A. Burnett, and P. Pasquier. 2012. Evaluating Musical Metacreation in a Live Performance Context. In *Proceedings of the Third International Conference on Computational Creativity (ICCC 2012)*. Dublin, Ireland, 140–144.

G. Fauconnier and M. Turner. 2008. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

J. C. Forth. 2012. *Cognitively-motivated geometric methods of pattern discovery and models of similarity in music*. Ph.D. Dissertation. Goldsmiths, University of London.

A. Friberg. 2004. A fuzzy analyzer of emotional expression in music performance and body motion. *Proceedings of Music and Music Science, Stockholm* 2005 (2004).

P. Gärdenfors. 2000. *Conceptual Spaces: the geometry of thought*. MIT Press, Cambridge, MA.

M. I. Garrido, J. M. Kilner, S. J. Kiebel, and K. J. Friston. 2007. Evoked brain responses are generated by feedback loops. *Proceedings of the National Academy of Sciences* 104, 52 (2007), 20961–20966.

W. D. Gorder. 1980. Divergent production abilities as constructs of musical creativity. *Journal of Research in Music Education* 28, 1 (1980), 34–42.

K. Grace and M. L. Maher. 2014. What to expect when youre expecting: the role of

unexpectedness in computationally evaluating creativity. In *Proceedings of the 4th International Conference on Computational Creativity, to appear*.
M. Grierson and C. Kiefer. 2011. Better Brain Interfacing for the Masses: Progress in Event-related Potential Detection Using Commercial Brain Computer Interfaces. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. ACM, New York, NY, USA, 1681–1686.
M. Grierson and C. Kiefer. 2014. Contemporary Approaches to Music BCI Using P300 Event Related Potentials. In *Guide to Brain-Computer Music Interfacing*, E. R. Miranda and J. Castet (Eds.). Springer London, 43–59.
J. P. Guilford. 1959. Traits of creativity. In *Creativity and its cultivation*, H. H. Anderson (Ed.). New York: Harper, 142–161.
J. P. Guilford. 1967. *The nature of human intelligence.* McGraw-Hill.
A. Hawryshkewich, P. Pasquier, and A. Eigenfeldt. 2010. Beatback : A real-time interactive percussion system for rhythmic practise and exploration. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2010)*. Sydney, Australia, 100–105.
M. Hayhoe and D. Ballard. 2005. Eye movements in natural behavior. *Trends in cognitive sciences* 9, 4 (2005), 188–194.
L. Hiller and L. Isaacson. 1959. *Experimental Music*. McGraw–Hill, New York.
L. Hiller and L. Isaacson. 1993. Musical composition with high-speed digital computers. In *Machine Models of Music*, S. M. Schwanauer and D. A. Levitt (Eds.). The MIT Press, Cambridge, MA, 9–22.
D. Hörnel and W. Menzel. 1998. Learning Musical Structure and Style with Neural Networks. *Journal of New Music Research* 22, 4 (1998), 44–62.
A. Horner and D. E. Goldberg. 1991. Genetic Algorithms and Computer-Assisted Music Composition. In *Proceedings of the International Computer Music Conference (ICMC 1991)*.
D. Huron and E. H. Margulis. 2010. Musical expectancy and thrills. In *Handbook of music and emotion: Theory, research, applications*, P. N. Juslin and J. A. Sloboda (Eds.). Oxford University Press.
D. B. Huron. 2006. *Sweet anticipation: Music and the psychology of expectation*. MIT press.
N. Jaušovec. 2000. Differences in cognitive processes between gifted, intelligent, creative, and average individuals while solving complex problems: an EEG study. *Intelligence* 28, 3 (2000), 213–237.
A. Jordanous. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4, 3 (2012), 246–279.
M. A. Kaliakatsos-Papakostas, A. Floros, and M. N. Vrahatis. 2012. Intelligent generation of rhythmic sequences using finite L-systems. In *Proceedings of the Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2012)*. 424–427.
J. Khatena and E. Torrance. 1976. *Manual for Khatena-torrance Creative Perception Inventory*. Technical Report. Stoelting Company.
K. H. Kim. 2006. Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity research journal* 18, 1 (2006), 3–14.
A. Koestler. 1964. *The act of creation.* Macmillan.
M. Kutas and S. A. Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307 (1984), 161–3.
K. Lange. 2009. Brain correlates of early auditory processing are attenuated by expectations for time and pitch. *Brain and cognition* 69, 1 (2009), 127–137.
C. León and P. Gervás. 2010. The role of evaluation-driven rejection in the successful

exploration of a conceptual space of stories. *Minds and Machines* 20, 4 (2010), 615–634.
A. Levisohn and P. Pasquier. 2008. BeatBender: subsumption architecture for autonomous rhythm generation. In *Proceedings of the International Conference in Advances on Computer Entertainment Technology (ACE 2008)*. 51–58.
C. Lew-Williams and J. R. Saffran. 2012. All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition* 122, 2 (2012), 241–246.
G. E. Lewis. 2000. Too Many Notes: Computers, Complexity and Culture in Voyager. *Leonardo Music Journal* 10 (01 Dec 2000), 33–39.
S. R. Livingstone, A. R. Brown, and R. Muhlberger. 2005. Playing With Affect: Music performance with awareness of score and audience. In *Australasian Computer Music Conference*, T. Opie and A. R. Brown (Eds.). Australasian Computer Music Association, Brisbane, Australia, 89–95.
S. R. Livingstone, R. Muhlberger, A. R. Brown, and W. F. Thompson. 2010. Changing Musical Emotion: A Computational Rule System for Modifying Score and Performance. *Computer Music Journal* 34, 1 (2010), 41–65. Volume 34, Number 1, Spring 2010.
S. R. Livingstone, W. F. Thompson, and F. A. Russo. 2009. Facial expressions and emotional singing: A study of perception and production with motion capture and electromyography. (2009).
J. London. 2012. *Hearing in time: psychological aspects of musical metre* (2nd ed.). Oxford University Press, Oxford, UK.
S. J. Luck, H. Heinze, G. Mangun, and S. A. Hillyard. 1990. Visual event-related potentials index focused attention within bilateral stimulus arrays. II. Functional dissociation of P1 and N1 components. *Electroencephalography and clinical neurophysiology* 75, 6 (1990), 528–542.
M. L. Maher. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*. Desire Network, 22–28.
C. Martindale, D. Hines, L. Mitchell, and E. Covello. 1984. EEG alpha asymmetry and creativity. *Personality and Individual Differences* 5, 1 (1984), 77–86.
C. Martindale, K. Moore, and J. Borkum. 1990. Aesthetic preference: Anomalous findings for Berlyne's psychobiological theory. *The American Journal of Psychology* (1990), 53–80.
J. B. Maxwell. 2014. *Generative Music, Cognitive Modelling, and Computer-Assisted Composition in MusiCog and ManuScore*. Ph.D. Dissertation. Simon Fraser University.
J. McCartney. 2002. Rethinking the Computer Music Language: SuperCollider. *Computer Music Journal* 26, 4 (2002), 61–68.
A. McLean and G. Wiggins. 2010. Tidal - Pattern Language for the Live Coding of Music. In *Proceedings of the 7th Sound and Music Computing Conference*.
S. Mednick. 1962. The associative basis of the creative process. *Psychological review* 69, 3 (1962), 220.
D. M. Mertens. 2014. *Research and Evaluation in Education and Psychology: Integrating Diversity With Quantitative, Qualitative, and Mixed Methods*. Sage Publications.
E. R. Miranda. 2006. Brain-computer music interface for composition and performance. *International Journal on Disability and Human Development* 5, 2 (2006), 119–126.
D. Moffat and M. Kelly. 2006. An investigation into people's bias against computational creativity in music composition. In *Proceeings of the International Joint Workshop on Computational Creativity*.

K. Monteith, B. Brown, D. Ventura, and T. Martinez. 2013. Automatic generation of music for inducing physiological response. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society (COGSCI 2013)*. 3098–3103.

K. Monteith, V. Francisco, T. Martinez, P. Gervás, and V. Dan. 2011. Automatic generation of emotionally-targeted soundtracks. In *Proceedings of the Second International Conference on Computational Creativity (ICCC 2011)*. México City, México, 60–62.

K. Monteith, T. Martinez, and D. Ventura. 2010. Automatic generation of music for inducing emotive response. In *Proceedings of the First International Conference on Computational Creativity (ICCC 2010)*. Lisbon, Portugal, 140–149.

A. North and D. Hargreaves. 2008. *The social and applied psychology of music*. Oxford University Press.

F. Pachet. 1990. Representing Knowledge Used by Jazz Musicians. In *Proceedings of the International Computer Music Conference*.

F. Pachet. 2000. Rhythms as emerging structures. In *Proceedings of the International Computer Music Conference (ICMC 2000)*. Berlin, Germany, 316–319.

F. Pachet. 2002. Playing with Virtual Musicians: the Continuator in practice. *IEEE Multimedia* 9, 3 (2002), 77–82.

F. Pachet. 2006. Enhancing individual creativity with interactive musical reflective systems. In *Musical creativity: multidisciplinary research in theory and practice*, I. Deliège and G. A. Wiggins (Eds.). Psychology Press, Hove, UK, Chapter 19, 359–375.

G. Papadopoulos and G. A. Wiggins. 1998. A Genetic Algorithm for the Generation of Jazz Melodies. In *Proceedings of STeP'98*. Jyväskylä, Finland.

P. Pasquier, A. Eigenfeldt, and O. Bown. 2014. Preface. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

C. V. Patricio and H. Honing. 2014. Generating expressive timing by combining rhythmic categories and Lindenmayer systems. In *Proceedings of the 50th Anniversary Convention of the AISB*, M. M. Al-Rifaie, J. Gow, and S. McGregor (Eds.). London, UK, 1–7.

M. T. Pearce. 2005. *The construction and evaluation of statistical models of melodic structure in music perception and composition*. Ph.D. Dissertation. City University London.

M. T. Pearce, M. H. Ruiz, S. Kapasi, G. A. Wiggins, and J. Bhattacharya. 2010. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage* 50, 1 (2010), 302–313.

M. T. Pearce and G. A. Wiggins. 2006. Expectation in melody: The influence of context and learning. (2006).

M. T. Pearce and G. A. Wiggins. 2007. Evaluating cognitive models of musical composition. In *Proceedings of the 4th international joint workshop on computational creativity*. Goldsmiths, University of London, 73–80.

M. T. Pearce and G. A. Wiggins. 2012. Auditory Expectation: The Information Dynamics of Music Perception and Cognition. *Topics in Cognitive Science* 4, 4 (2012), 625–652.

S. Phon-Amnuaisuk, A. Smaill, and G. Wiggins. 2006. Chorale harmonization: A view from a search control perspective. *Journal of New Music Research* 35, 4 (2006), 279–305.

S. Phon-Amnuaisuk, A. Tuson, and G. A. Wiggins. 1999. Evolving Musical Harmonisation. In *Proceedings of ICANNGA'99*. Springer, Portorož, Slovenia, 229–234.

C. C. Preston and A. M. Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica* 104, 1 (2000), 1–15.

M. Puckette. 1996. Pure Data. In *Proceedings of the International Computer Music Conference*. International Computer Music Association, San Francisco, 224–227.

G. Ritchie. 2001. Assessing creativity. In *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science*. Citeseer.
G. Ritchie. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17, 1 (2007), 67–99.
D. Rosenboom. 1990. The performing brain. *Computer Music Journal* (1990), 48–66.
R. Rowe. 2001. *Machine Musicianship*. MIT Press, Cambridge, MA, US.
R. Rowe. 2008. Algorithms and Interactive Music Systems. *Contemporary Music Review* (2008).
S. Rubin and M. Agrawala. 2014. Generating emotionally relevant musical scores for audio stories. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 439–448.
J. R. Saffran, R. N. Aslin, and E. L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274, 5294 (1996), 1926–1928.
R. Saunders. 2012. Towards Autonomous Creative Systems: A Computational Approach. *Cognitive Computation* 4, 3 (2012), 216–225.
R. Saunders and J. S. Gero. 2001. Artificial creativity: A synthetic approach to the study of creative behaviour. *Computational and Cognitive Models of Creative Design V, Key Centre of Design Computing and Cognition, University of Sydney, Sydney* (2001), 113–139.
J. Schmidhuber. 2010. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *Autonomous Mental Development, IEEE Transactions on* 2, 3 (sept. 2010), 230–247.
N. Schwarz. 1999. Self-reports: how the questions shape the answers. *American psychologist* 54, 2 (1999), 93.
R. N. Shepard. 1962a. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* 27, 2 (1962), 125–140.
R. N. Shepard. 1962b. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* 27, 3 (1962), 219–246.
R. T. Solberg. 2014. Exploring the club experience: Affective and bodily experiences of electronic dance music. In *International Conference of Students of Systematic Musicology*.
L. Steck and P. Machotka. 1975. Preference for musical complexity: Effects of context. *Journal of Experimental Psychology: Human Perception and Performance* 1, 2 (1975), 170.
M. J. Steedman. 1996. The Blues and the Abstract Truth: Music and Mental Models. In *Mental Models In Cognitive Science*. Erlbaum, Mahwah, NJ, 305–318.
J. Stockholm and P. Pasquier. 2008. Eavesdropping: Audience Interaction in Networked Audio Performance. In *Proceeding of the 16th ACM international conference on Multimedia - MM '08*. ACM Press, New York, New York, USA, 559–568.
J. Stockholm and P. Pasquier. 2009. Reinforcement Learning of Listener Response for Mood Classification of Audio. In *2009 International Conference on Computational Science and Engineering*, Vol. 4. IEEE, 849–853.
T. Stoll. 2014. Genomic: Combining Genetic Algorithms and Corpora to Evolve Sound Treatments. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.
D. Temperley and D. Sleator. 1999. Modeling Meter and Harmony: A Preference-Rule Approach. *Computer Music Journal* 23, 1 (01 Mar 1999), 10–27.
E. Torrance. 1998. *The Torrance tests of creative thinking—technical manual figural forms A & B*. Technical Report. Scholastic Testing Service, Inc., Bensenville, IL, US.
M. Tunner and G. Fauconnier. 1995. Conceptual integration and formal expression. *Metaphor and Symbol* 10, 3 (1995), 183–204.
A. Turing. 1950. Computing machinery and intelligence. *Mind* LIX, 236 (1950), 433–

60.
D. Tzimeas and E. Mangina. 2006. Jazz Sebastian Bach: A GA System for Music Style Modification. In *Proceedings of the International Conference on Systems and Networks Communication (ICSNC '06)*. IEEE Computer Society, Washington, DC, USA, 36–42.
D. Tzimeas and E. Mangina. 2009. Dynamic Techniques for Genetic Algorithm-Based Music Systems. *Computer Music Journal* 33, 3 (2009), 45–60. Volume 33, Number 3, Fall 2009.
T. Veale, P. Gervás, and R. P. y Pérez. 2010. Computational creativity: A continuing journey. *Minds and Machines* 20, 4 (2010), 483–487.
J. N. Vold. 1986. *A study of musical problem solving behavior in kindergarten children and a comparison with other aspects of creative behavior*. Ph.D. Dissertation. University of Alabama.
M. A. Wallach and N. Kogan. 1965. *Modes of thinking in young children.* New York.
G. Wallas. 1926. The art of thought. (1926).
I. Wallis, T. Ingalls, and E. Campana. 2008. Computer-Generating emotional music: The design of an affective music algorithm. *DAFx-08, Espoo, Finland* (2008), 7–12.
G. Wang, P. R. Cook, and S. Salazar. 2015. ChucK: A Strongly Timed Computer Music Language. *Computer Music Journal* 39, 4 (2015), 10–29.
Q. Wang, P. Cavanagh, and M. Green. 1994. Familiarity and pop-out in visual search. *Perception & psychophysics* 56, 5 (1994), 495–500.
P. R. Webster. 1987. Conceptual bases for creative thinking in music. In *Music and child development*. Springer, 158–174.
G. A. Wiggins. 2006a. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19, 7 (2006), 449–458.
G. A. Wiggins. 2006b. Searching for Computational Creativity. *New Generation Computing* 24, 3 (2006), 209–222.
G. A. Wiggins. 2007. Review article: 'Computer Models of Musical Creativity' by David Cope. *Literary and Linguistic Computing* 23, 1 (2007), 109–116.
G. A. Wiggins. 2009. Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. In *Proceedings of 11th IEEE International Symposium on Multimedia*. 477–482.
G. A. Wiggins. 2012a. The future of (mathematical) music theory. *Journal of Mathematics and Music* 6, 2 (2012), 135–144.
G. A. Wiggins. 2012b. The Mind's Chorus: Creativity before Consciousness. *Cognitive Computation* 4, 3 (2012), 306–319.
G. A. Wiggins. 2012c. Music, Mind and Mathematics: Theory, Reality and Formality. *Journal of Mathematics and Music* 6, 2 (2012), 111–123.
G. A. Wiggins. 2012d. On the correctness of imprecision and the existential fallacy of absolute music. *Journal of Mathematics and Music* 6, 2 (2012), 93–101.
G. A. Wiggins and J. Forth. 2016. Computational Creativity and Live Algorithms. In *The Oxford Handbook of Algorithmic Music*, R. T. Dean and A. McLean (Eds.). Oxford University Press. In preparation.
G. A. Wiggins, M. T. Pearce, and D. Müllensiefen. 2009. Computational Modelling of Music Cognition and Musical Creativity. In *The Oxford Handbook of Computer Music*, R. Dean (Ed.). Oxford University Press, Oxford, UK, 383–420.
G. A. Wiggins, P. Tyack, C. Scharff, and M. Rohrmeier. 2015. The Evolutionary Roots of Creativity: mechanisms and motivations. *Philosophical Transactions of the Royal Society B: Biological Sciences* Issue on Musicality, in press (February 2015).
M. S. Worden, J. J. Foxe, N. Wang, and G. V. Simpson. 2000. Anticipatory Biasing of Visuospatial Attention Indexed by Retinotopically Specific-Band Electroencephalography Increases over Occipital Cortex. (2000).

M. Wright and D. Wessel. 1998. An Improvisation Environment for Generating Rhythmic Structures Based on North Indian "Tal" Patterns. In *International Computer Music Conference*. International Computer Music Association, Ann Arbor, Michigan, 125–128.

G. N. Yannakakis and H. P. Martínez. 2015. Ratings are Overrated! *Frontiers in ICT* 2, 13 (2015).

Received May 2015; revised Dec 2015; accepted XXX 2016