Validation of Harmonic Progression Generator Using Classical Music

Adam Burnett
Cognitive Science
Simon Fraser University
Burnaby, BC Canada
ajb14@sfu.ca

Evon Khor
Cognitive Science
Simon Fraser University
Burnaby, BC Canada
ewk@sfu.ca

Philippe Pasquier
Interactive Arts and Technology
Simon Fraser University
Surrey, BC Canada
pasquier@sfu.ca

Arne Eigenfeldt
Contemporary Arts
Simon Fraser University
Vancouver, BC Canada
arne_e@sfu.ca

Abstract

We evaluate the output of a Markov model-based harmonic progression generator, a classic model for corpus-based computational creativity. 87 participants performed a discrimination task classifying 20 musical excerpts as either human-composed or computer-composed. Also recorded was each participant's level of confidence in their choice. Results indicated that while overall performance was above what would be expected from random guessing, further analysis revealed this was due to the human-composed pieces being much easier to identify than computer-composed pieces. Assessed separately, participants were unable to identify computer-composed pieces above chance-levels. We suggest improvements to the experimental design that could be implement in future evaluations.

Introduction

The trouble with evaluating artistic creativity is that it is difficult to establish objective criteria with which to judge the resulting creative artifact. This problem is compounded when the source of the artifact is *itself* a creative software. Computational creativity results in creations of creations, or *metacreations* (Whitelaw 2004), that differ from the artifacts we are used to encountering. As they are produced by machines that vary in their level of autonomy and in the amount of user-interaction they require in order to function, we must keep in mind a different set of considerations when we evaluate the resulting pieces.

The evaluation of aesthetics in metacreations is a fixture in the computational creativity literature (Eigenfeldt and Pasquier 2010; Eigenfeldt and Pasquier 2011; Pease, Winterstein, and Colton 2001; Stamp, Isenberg, and Carpendale 2007; Wallraven, Cunningham, Rigau, Feixas, and Sbert 2009). Whereas grading intelligence is a straight-forward matter of assessing how closely or quickly an individual can reach the optimum solution to a formally specifiable problem, there is usually no clear "goal" or "problem" that needs to be solved in a creative artwork; it often exists for its own sake, and to be enjoyed. When judging creative works such as music, it often comes down to relying on the subjective impressions of experts in the relevant domain (music critics, musicologists) or by quantifying subjective impressions in some measurable way (album sales, concert attendance).

Relying on entirely subjective measures is undesirable because it is not sufficient to simply test whether human listeners find computer-generated music creative or enjoyable or emotional: the mind is capable of finding patterns, design, and intention in random noise, deriving pleasure in the beauty of living organisms and ecosystems which were "designed" by the unguided, unintelligent processes of evolution, and sometimes even in randomness itself. To fairly and accurately assess the quality of computer-generated music, we must devise some sort of objective means to do so, even if that means indirectly measuring an effect of that creativity rather than directly measuring the creativity itself.

In the following we will describe previous attempts to evaluate computer-generated music, and then present our evaluation of the harmonic progression generator developed by Arne Eigenfeldt and Philippe Pasquier (2010).

Background

A problem is inevitably encountered when one tries to merge a scientific discipline concerned with objectivity, like Artificial Intelligence, with the subjectivity inherent in the creation, appreciation, and evaluation of art. This is the challenge for anyone proposing ways to evaluate machine creativity. As noted by Spector and Alpern (1994), there is no universally agreed-upon theory of aesthetic value within the artistic community. How then do we know when we have a *computational* artist? Approaches to this problem generally fall into two main camps. The first advocates a reliance on human judgements, particularly of the art-world and museum-going public, by holding art shows and getting feedback. However, this requires a lot of time and resources and is not always practical nor reliable. The other approach recommends the creation and application of codified, formalized evaluation criteria with which to judge a computational artist's creations. This method is especially popular in computer music evaluation as many forms of music can be formulated to follow a rule-system. There are three problems with this, however: 1) many existing formulations are "dead forms", which would penalize works in more contemporary genres for which detailed formalizations have not yet been established, 2) it is not evident that strict adherence to rules of a particular art-form or genre indicate aesthetic value. Meeting this criterion might indicate nothing more than aesthetically mediocre and boring, formulaic work, and 3) many formulations are rigid and

once established may not lend themselves to generalization across genres, essentially punishing novel works for being too original, even if they are of high quality.

Alternatively, Colton (2008) suggests that, rather than a focus on the input and output, *how* a creative work is produced is critical to its being perceived as creative. Colton asks us to consider the question of whether we label works as "creative" based on their quality, or whether we determine the quality of works based on how creative we found the process that generated them to be. Colton notes that in painting audiences are concerned with the process that led to the final product, and that this affects their enjoyment of the piece. In fact, it is noted that often the actual aesthetic quality of an artwork has little to do with how creative the work is perceived to be (consider Duchamp's *Fountain*, which was nothing more than a urinal). One conundrum which follows from this approach is that when *too little* is known about the process, we cannot evaluate whether or not it is creative, and if we know *too much* about the process, it is regarded as too mechanical and deterministic, leaving no room left for "creativity" to be exercised.

Colton presents a model of art appreciation, proposing that there are three judgements that consumers make about the creative process when determining how much they like a piece. These are: 1) the perceived effort required during the process, 2) the ingenuity of the process, and 3) the skill needed to carry out the process. From these Colton derived *The Creative Tripod*, which defines the three properties a system must possess in order to be judged as being creative: *skill*, *appreciation*, and *imagination*. Without skill, nothing can be produced; without appreciation, nothing of *value* can be produced, and without imagination, nothing *original* can be produced. The tripod analogy highlights the need for all three properties to be present in order of the label of "creative" to stand.

Whereas Colton directs attention on the process, Ritchie (2007) de-emphasizes the internal processes and favours focusing solely on the output. Ritchie argues that when assessing creative artifacts, we should be faithful to the traditional use of the word "creativity", which is tied to subjective human judgements. This, combined with the stance that we should only evaluate observable creative behaviours, levels the playing field and allows us to assess both human-produced and machine-produced creative works fairly and without bias. Ritchie warns that considering *both* the artifact and the process would introduce a fatal circularity: we would be left arguing that an artifact is creative because the process that produced it was creative, and that we know the process that produced it was creative because the artifact it produced was creative.

Discrimation Tasks

Pearce, Meredith, and Wiggins (2002) define four motivations for developing generative music systems: 1) to implement them as tools for personal use and/or 2) for general compositional use, 3) to provide theories of musical style, and 4) to provide cognitive theories of processes in compositional expertise. These four motivations can be collapsed into two general categories, the first of which is to use generative music systems as creative tools to produce original music, the other is to use these systems as a way to model theories of musical style and cognition. We will not be discussing the latter category any further here.

The problem of evaluating creativity mirrors a similar problem that befell early artificial intelligence researchers: *how do we evaluate machine intelligence*? It was difficult to say whether or not a machine could ever be said to think or demonstrate intelligence because there was little agreement on what those words would mean in the context. Today we face the same problem with machine creativity, unable to unanimously agree on what is meant by "creativity" in the question: *how do we evaluate machine creativity?*

Alan Turing (1950) famously suggested a way to tackle the problem. He had us consider a party game (the "imitation game") where a judge tries to determine which of two unseen players is pretending to be a woman; it is the job of the man to fool the judge by responding in the way he thinks a woman would, and it is the job of the woman to assist the judge in exposing him. With that in mind, Turing suggest that instead of trying to answer the impossible question of *can machines think?*, we should reformulate the question into something we can answer: *are there imaginable digital computers which would do well in the imitation game?* That is, could a computer program ever be designed that could successfully convince a judge that he was conversing with a human? This hypothetical procedure came to be known as the *Turing Test*. How this approach might be adapted for the problem of machine creativity is apparent: reformulate the question from whether or not a composition system is creative, and instead ask whether it does well in the "imitation game".

A popular method of evaluating generative music systems is to run a Turing-style test on the system's output (Boden 2010). This involves comparing computer-generated compositions to human-generated compositions through participant evaluation of the various pieces. Ariza (2009) cautions against the use of term "musical Turing Test" since intelligence of a generative music system cannot be determined by evaluating its output. The Turing Test has underlying assumptions on which it builds its criteria for machine intelligence: humans have minds, and natural language is sufficient to represent the mind; thus, if a machine is indistinguishable from a human in discourse, then it too has a mind. Joseph Weizenbaum's ELIZA is an early example of a system suited for the Turing Test. The ELIZA system was able to fool human interrogators; however, can we say that ELIZA is intelligent? John Searle's Chinese Room Argument suggests that a system is able to fool its interrogators without knowing anything about what it is doing; and so, having a façade of intelligence does not mean that the system is actually doing anything intelligent.

To test the outputs of generative music systems, it is possible to tweak the criteria of the Turing Test to accommodate the evaluation of musical outputs. Instead of having a text-based medium, sound symbols or forms are used. The

interrogator is replaced by a critic who may or may not interact with the system. Harnad (2000) labels tests of these sorts as toy Tests (tTs) instead of Turing Tests (TTs). In the Musical Output toy Test (MOtT), the critic is presented with musical pieces from two composer-agents. One of these agents is human, and the other, of course, is a machine. Based only on these works, the critic must attempt to distinguish the human from the machine.

Caution must be taken when interpreting the results of this type of test. For one, what criteria the listener uses when trying to discern between the machine- and human-composed pieces needs to be asked explicitly, and even so, listeners may fail to even be cognisant of their decision-making processes. Second, musical judgements are influenced by any combination of factors and can vary greatly from individual to individual. Furthermore, it is important to keep in mind that these tests are surveys of musical judgement and not of whether the system has thought or intelligence.

Boden (2010) cites David Cope's *Experiments in Musical Intelligence* (EMI) system, which generated music in the style of music contained in a supplied database, and notes how those listeners with some musical experience had difficulty determining whether the pieces it produced were human-composed or not. However, those with more extensive familiarity with, for example, Mozart were more readily able to distinguish true Mozart-composed pieces from the EMI-composed Mozart-esque pieces, though they there were unable to tell the difference between the EMI-composed pieces and human-composed pieces which were both meant to *mimic* Mozart's style. Even when EMI failed to perfectly mimic the intended style, it still was able to produce pieces that demonstrate proper compositional technique. Performance in a Turing Test thus varies heavily depending on exactly *what* is being tested (ability to mimic a particular composer? Ability to follow compositional conventions? Ability to produce interesting melodies?).

Another obstacle for Turing Tests is that they require the cooperation of the human judges. People have been known to retract their praise for computer-generated works upon learning of their synthetic source, protesting that it is requisite of art to have been produced by a human being, possessing all the facilities that enable one to express and communicate human emotion and experience. Some have refused to even give audience to a creative work knowing that it was produced by a machine, as happened to David Cope when debuting EMI. The reason cited is the belief that art requires creativity, and the belief that computers cannot be creative precludes computers from creating art. This prejudice will prevent some from ever accepting the results of a Turing Test, even if it is deemed internally successful.

Pearce and Wiggins (2001) proposed an objective framework for evaluating computer-generated musical compositions which, as they themselves point out, elicits comparisons to the Turing Test. The framework was developed in response to problems they identified with previous attempts to evaluate music composed by computer systems. They distinguished two kinds of evaluation: the *critic* and what we could call the *evaluation-proper*. The critic is part of the music-generating system itself and helps guide the development of the composition by evaluating the intermediate products of the system. The evaluation-proper is that which we are mainly concerned with here, and is unfortunately the more elusive of the two: it is the process and methods of establishing whether the compositions produced by the system satisfy the specified conditions of creativity.

Pearce and Wiggins highlight the necessity of objective measurements when evaluating machine creativity, as an objective approach to evaluation would be consistent with standard scientific investigation. Empirical science carries a respectable weight, and if a creative system could survive the sort of rigorous testing expected in scientific domains, then the results would be far more compelling than the wishy-washy, subjective evaluations seen elsewhere.

The existing evaluation methods Pearce and Wiggins reviewed are criticized for failing to confirm to these standards of objectivity, in part due to the presence of programmers' bias in the critic algorithms embedded in a number of the systems they discussed. They also note that subjective impressions are very imprecise and potentially unreliable: it is difficult to ensure that a group of human evaluators are following the same criteria.

In reaction to these shortcomings, Pearce and Wiggins layout a method of evaluating composition systems that maintain objective integrity. In order to be objective, a specific compositional aim must be explicitly established beforehand—if the goal is to mimic the style of a specific Baroque composer, the system should not receive a positive evaluation score because it is able to generate a realistic progression of 20th century jazz chords. To eliminate programmer bias, the critic should be derived from a pattern extracted from a data set of existing compositions using a machine learning algorithm. Once music is composed that is able to satisfy the critic, an evaluation using human participants is conducted. The participants should be played both music composed by the system and from the data set itself, and then tested to see if they are able to distinguish the two.

Having participants simply indicate whether they think a piece of music is computer-composed or not frees us from the subjective question of whether the computer-generated work is creative, enjoyable, or emotional, and instead allows us to home in on the *objective fact* about whether or not humans are able to tell whether the works are computer composed. Reformulating the question from *can this system produce creative works?* to *can this system produce works indistinguishable from the human composers in the data set?* creates a predictable, testable, and perhaps most importantly, a *clearly refutable* claim, as would be expected from a rigorous scientific experiment.

Experiment

In the framework of Pearce and Wiggins (2001), no at-

tempt was made to deceive the judges/participants about the nature of the experiment: participants were explicitly informed that they were comparing human- and machine-composed music. Along with this candid approach, our experiment resembled the framework described by Pearce and Wiggins in many additional ways. Our experiment differs however in that we compare the performance of both musicians and nonmusicians, and go beyond offering a simple binary choice between machine-composed and human-composed by providing a 4-point scale which will reflect both the participants' choice and their confidence in their choice.

In our study, we aim to evaluate the quality and particularly the robustness of the harmonic progression generator developed by Arne Eigenfeldt and Philippe Pasquier (2010). We sought to determine how successful the program is at generating harmonic progression of the same quality and style as human composers from traditional classical style periods. To do this, we had two groups of human participants, varying in their musical fluency, attempt to distinguish musical excerpts generated by the program from those written by human composers.

System Description

The system we are evaluating uses a third-order Markov model to derive harmonic progressions from a supplied corpus (Eigenfeldt and Pasquier 2010). This allows versatility as the particular rules from a style-period or genre do not have to be hard-coded into the program. Instead, the appearance of the rules emerge from the reliance on the corpus to guide the generation of progressions. By foregoing set rules, the system is not biased toward only producing progressions that follow traditional harmonic and voice leading conventions, but can just as competently function within the harmonic freedom of 20th century music if provided with a sufficiently rich corpus.

In contrast to many other music generators, the system was designed to function and respond to user request. The user is given the ability to specify the number of bars to be generated, a target bass line, the level and variation of harmonic complexity, and the voice-leading tension of the generated chords. These vectors help select the best candidate among the generated Markov conditional probability distributions of chord transitions. The system is written in MaxMSP and is available on its first author's webpage[1]. A full outline of the system is provided in Eigenfeldt and Pasquier (2010).

The corpus which serves as input to the system consists of pre-processed MIDI files: all musical content is reduced to a sequences of chords (and their durations) with controller data indicating the beginning of phrases and cadences.

Participants

The participants were recruited from Simon Fraser University and the University of British Columbia. Participation was incentivized by offering four $50 prizes to be distributed upon completion of the study.

To increase the resolution of our test of the harmonic progression generator, we compared the performance of two independent groups: musicians and nonmusicians. Much like a spoken language, well-written music is constrained by and emerges from conventions and rules and patterns. If one were to do a validation of a spoken or written language-generating program using human participants as judges, it would clearly be necessary to have the participants be fluent in the target language. It is for this reason that we found that in order to perform an accurate validation, it was critical to test the difference between musicians and nonmusicians in this task.

For the purpose of this experiment, only those with formal training in classical musical analysis were deemed "musicians"—mere proficiency with an instrument did not suffice. While instrumentalists are indisputably "musicians", we were exclusively interested in those students who have spent time studying and analyzing classical music scores and may have developed an ability to identify unusual harmonic choices and other errors that might arise in a machine-generated composition. Therefore, we decided that the musician group would consist of students who have received two or more years of classical musical training at a post-secondary institution. To ensure sufficient group-size, we also admitted those who have received at least 5th grade certification in the royal conservatory of music (or equivalent). The group consisting of laypeople (nonmusicians) was screened during the survey to ensure their musical naïvety.

Music Selection

Corpus For our study, we used harmonic progressions derived from a corpus of classical music (a mixture of Classical and Romantic style periods). Only chords already present in the pieces that made up the corpus found their way into the generated excerpts. We presented ten excerpts of music generated by the system and ten from classical pieces adapted to match the non-melodic presentation style of the computer-generated pieces.

The following is a list of the pieces that made up the corpus from which the computer-generated pieces were derived. The corpus is divided into five sections, each containing five to six pieces from the same composer to ensure consistency of style. We have tried to maintain consistence of form in our selections as well. Two of the pieces from each section were included in the survey as the human-composed pieces (marked in bold), and two progressions were generated from each section using the harmonic progression generator. As we are evaluating the quality of progressions produced by the system rather than determining the limits of its functionality, the setup described here corresponds to a typical use of the system.

Frédéric Chopin: **Nocturne in Eb Major Op. 9, No. 2**; Nocturne in F# Major Op. 15, No. 2; Nocturne in G minor, Op. 15, No. 3; Nocturne in Db Major, Op. 27, No. 2; **Nocturne in F major Op. 55, No. 1**.

[1] http://www.sfu.ca/~eigenfel/arne/main.html

Antonín Dvoř ák: **Humoresque**, Legend, **Slavonic Dance No. 1**, Slavonic Dance No. 2, Symphony No. 9 "From The New World" Second Movement, Valse Gracieuse.

Johannes Brahms: Symphony No. 1 In C Minor 3rd Movement, **Symphony No. 2 In D 3rd Movement**, **Symphony No. 3 in F 2nd Movement**, Symphony No. 3 in F 3rd Movement, Symphony No. 4 In E minor 3rd Movement, Hungarian Dance No. 5.

Felix Mendelssohn: **Consolation**, If With All Your Hearts, Spinning Song, O Rest In The Lord, **Scherzo in E Minor**, Venetian Boating Song (from Songs Without Words).

Robert Schumann: About Strange Lands And People, Träumerei, (from Scenes from Childhood), **The Happy Farmer** (from Album for the Young), **Piano Concerto in A Minor**, The Wild Horseman, Arabesque.

Processing The original Turing Test was not an assessment of a machine's ability to mimic speech, and neither was our experiment a test of the system's ability to creatively interpret and audibly produce music like a performer, but merely to compose it. Therefore, all pieces used in the experiment were "performed" and recorded using Kontakt Player (Native Instruments) and Cakewalk Sonar 4 (Cakewalk).

The system we evaluated requires pre-processing of the items in the data set: as the system is concerned only with analyzing and generating harmonic progressions, it was designed to receive as input MIDI files that conform to a specific format of block chords in closed position with the bass note separately specified. In the name of efficiency, rather than manually analyzing the harmony in our chosen classical pieces, we utilized "The Real Little Classical Fake Book" (Hal Leonard Corp. 1993), a large collection of classical themes transcribed for piano, and simply discard the melodic line and sequenced the harmonies and harmonic rhythms into MIDI files using the chord symbol realization plugin (which generates notes from chord symbols) for the Sibelius scorewriting software (Sibelius).

As we planned to test both musicians as well as nonmusicians, we recognized it was important that the human-composed pieces we chose be unfamiliar enough to reduce the likelihood that either groups would recognize their harmonic structure. Though we imagine that the elimination of the melodic line alone sufficiently obscured the identity of the pieces (as will transposition to a different key and changing the tempo), discretion was taken to ensure that pieces that obtain most of their notoriety from their harmonic sequences were excluded.

Determining which computer-generated pieces would make it into the final test was done by selecting those that most closely conformed to a pre-specified criteria. To ensure the feasibility of the study, it was decided to restrict the length of the harmonic sequences generated to around 8-bars and ensure that the progression ended on the tonic or dominant chord (regardless of the chord that preceded it), or in a cadence. The first four bars of one of the computer-generated pieces is presented in Figure 1. Note that the system encodes rhythm and can generate more than one chord per bar.

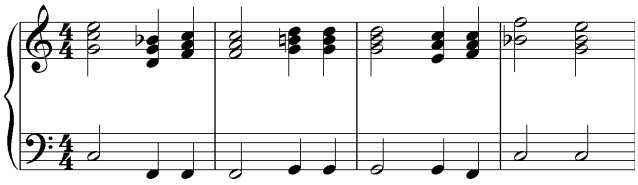

Figure 1. four-bar example of a "Brahms-inspired" excerpt.

Procedure

87 participants, a mix of students and faculty from Simon Fraser University and the University of British Columbia, were provided a URL to an online survey[1]. The survey was built using Drupal (Drupal) and a number of modules to enable audio playback and time tracking (to ensure that the participants were listening to the musical excerpts in full at least once). Participants were presented with a consent page indicating that their consent would be assumed by their completing the survey. They were then directed to a screen inquiring about their musical training (to enable us to assign them to the correct experimental group), followed by instructions detailing how to complete the survey. The instructions were upfront about the purpose and methods of the experiment; participants were informed that they would hear a mix of human-composed music and machine-composed music. No deceptive protocols were utilized.

Participants were presented one piece of music at a time, presented in a pre-established random order (limitations of the implementation prevented us from having each participant experience a unique ordering of questions). After listening, they were asked to rate the likely composer of each piece on a 4-point scale with 1 being "definitely human", 2: "probably human", 3: "probably computer" and 4: "definitely computer".

We decided to avoid using an odd-numbered scale for two reasons: we first wanted to discourage participants from disengaging from the task and choosing a neutral rating of 3 throughout. If participants lack certainty, this will be reflected in a greater proportion of "probably X" responses. We wanted to encourage participants to provide their best guess instead of defaulting to a safe "don't know" choice as it has been shown in other perceptual judgement tasks that participants underestimate their ability and that when forced to make a choice they often choose the appropriate response (Brown 1910). Gilljam and Granberg (1993) suggest that the presence of "don't know" options in questionnaires might encourage even those with definite opinions to chose the more cautious response. Poe et al. (1988) found that, when concerned with question testing factual knowledge, there is little difference in the responses

[1]The survey can be accessed at the following URL: http://magnum-interactive.com/metacreation

on questionnaires with and without a "don't know" option, but that excluding a neutral choice resulted in more usable data. A study by Alwin and Krosnick (1991) also found that including a "don't know" option did not improve the reliability of the results.

Secondly, the 4-point scale allowed us to reserve the ability to collapse the data into a binary human/computer choice, as well as compare the frequencies of "definitely X" to "probably X" selections between musicians and non-musicians later on for statistical analysis.

Following the questionnaire, participants were thanked for their participation and asked to indicate whether or not they recognized any of the progressions (and specify what they thought they were if they did), and what strategies (if any) they used to determine if the pieces were human- or computer-composed. Participants were then directed to a separate website where they provided their email address. Here they could indicate whether they wanted to be contacted about the results of the experience and/or be entered into the prize draw. As this section is separate from the survey-proper, it prevented us from matching survey answers to identifiable e-mail address, preserving anonymity.

We hypothesized that if the harmonic progression generator is capable of creating music of a quality and style similar to human composers, we should expect to see the performance of the two groups be similar to that which would result from random guessing (null hypothesis). If there *is* a detectable difference between the computer-generated and human-composed originals (that is: it is possible to distinguish the two), we should expect to see the nonmusicians perform either close to or slightly-above chance levels, and the musicians out-perform the nonmusicians with a performance even further from chance levels in the direction of correct classification.

We used one sample t-tests to compare musicians and nonmusicians to chance levels, and two sample t-tests to compare the mean scores of the groups. Tests were conducted using Bonferroni corrected alpha level of 0.005 (0.05/10). Comparisons were also made between these two groups' confidence with their choices as derived from their proportions of 1s and 4s compared to 2s and 3s on the 4-point scale.

Analysis of Data

Performance Participants were given four options when indicating their level of musical experience. They could specify that they had at least 2 years of a Bachelor's degree in music (Bachelor's), had achieved 5th grade certification in the royal conservatory of music or equivalent (Royal Cons.), had some unspecified formal musical training (Some), or no training (None). Our original "musician" category collapsed the data from the Bachelor's and Royal Cons. groups together, while the "nonmusicians" are made up of participants from the Some and None group. Table 1 shows the different groups' overall performance on the discrimination task.

Experience	*mean*	*t-score*	*p*	*df*
musicians	11.92 (3.27)	2.633	0.0164	19
nonmusicians	11.62 (2.46)	5.386	< .0001	66

Table 1. mean of correct answers out of 20, t-score (compared to chance), significance level, and degrees of freedom. Note. Standard deviations appear in paratheses.

As a test of our first hypothesis, one sample t-tests were used to compare performance to chance-levels (given that the questions were binary, 10 good guesses out of 20 is the mean for chance level). The results were not as anticipated: nonmusicians did significantly better than chance, leaving us unable to retain the null hypothesis (that the quality or style of the computer-generated pieces are indistinguishable from the quality and style of the human-composed pieces).

These data appear at first glance to be in the opposite directions of what we expected. Further analysis however revealed that by only looking at participants' total scores we had overlooked an interesting pattern buried in the data. Inspired by a comparable analysis conducted in Pearce and Wiggins (2001), when scores on identifying human-composed pieces were analyzed independently from scores identifying computer-composed pieces, a much different picture of the results emerged. Table 2 shows the results of this analysis.

Experience	*mean*	*t-score*	*p*	*df*
musicians (H)	6.550 (1.82)	3.808	0.0012	19
musicians (C)	5.300 (1.92)	0.698	0.4936	19
nonmusicians (H)	6.477 (1.51)	8.003	< .0001	66
nonmusicians (C)	5.089 (1.99)	0.369	0.7136	66

Table 2. results broken down by group and compositional source. (H) = human-composed. (C) = computer-composed.

When tested with a one sample t-test, this analysis shows that while participants were able to classify human-composed pieces (H) well above chance-levels (5 out of 10), their performance identifying computer-composed pieces (C) was *not* significantly different from chance-levels. Human-composed pieces were much more easily identifiable as human-composed than computer-composed pieces were identifiable as computer-composed. However, we still failed to see any statistically significant difference between the scores of the musician and nonmusician groups.

Confidence We also measure the level of confidence the

participants experienced for each question in the discrimination task. Confidence for each question was determined by assigning two points for "definitely" answers and one point for "probably" answers. A percentage was calculated using the score and the maximum possible score (thus a score of 100 would mean that the participant gave a "definitely" answer on every question). Average group confidence scores are indicated in Table 3.

Experience	*mean*	*n*
musicians (H)	67.25 (11.06)	20
musicians (C)	64.00 (12.84)	20
nonmusicians (H)	67.46 (12.83)	67
nonmusicians (C)	63.06 (11.18)	67

Table 3. confidence scores by group and compositional source.

While comparisons between groups' confidence are not statistically significant (there was no difference in confidence between experts and laymen), if the group means do in fact hint at a general tendency, they would indicate that participants are more confident about their answers when classifying human-composed pieces. This would be consistent with the analysis of performance that indicates that participants are likely to correctly identify these pieces as human-composed.

Strategies In the written response section of the survey, out of 87 participants, and out of only three who ventured guesses, only one correctly identified that they heard a progression taken from a Chopin piece (though they did not specific the piece's name). For the rest, participants seemed to rely on a number of factors to help them correctly identify the pieces' compositional source. Participants classifying human-composed pieces indicated that they listen for qualities such as *depth, clarity, complexity, feeling, life, regularity in rhythm, consonance, variety of dynamics, fluidity, subtly, repetition, pleasantness, simplicity*, and *logic*. When trying to identify computer-composed pieces, participants indicated that they listened for *repetition, simplicity, increased speed, odd resolutions, invariable rhythm, dissonance, lack of feeling, symmetry, rigidity, formality, awkwardness, logic, choppiness, static dynamics*, and *disorganization*. Interestingly, a number of these properties overlap: participants trying to identify both human and computer-composed pieces claimed to be listening for *simplicity* and *logic*, and participants within each condition often were looking for opposite properties to help identify the same source.

Conclusion

Participants listened to a series of 20 harmonic progressions and indicated whether they thought each was human- or computer-generated, along with a rating of their confidence for each choice. We hypothesized that participants would not perform significantly better than chance at this task.

Overall, participants *did* discriminate between the human-composed material and the progressions generated by the system. However, examining the results in more detail revealed something unexpected. When looking at participant responses to trials containing computer-composed progressions in isolation, it was found that participants were not capable of identifying the pieces generated by the harmonic progression system as computer-composed. Surprisingly, participants were nonetheless capable of identifying the human-composed pieces above chance levels. This results suggest that humans have a "natural" tendency to correctly recognize human-generated content. This would explain while our validation test failed. Further study would be needed to generalize this last finding. We believe that this tendency could be of interest to the computational creativity community as well as for cognitive sciences in general.

There are a number of changes to our experimental design that would be worth attempting in follow-up studies. The group sizes in the present experiment were quite heterogeneous, and the results seem to suggest that a larger number of participants qualifying for the Bachelor's group could provide us with valuable data.

We would also likely benefit from randomizing the presentation order of the musical excerpts or offering a more extensive "practice" section in future experiments. The collective results of all participants, plotted against time, gave a Pearson's correlation of $r = 0.54$, suggesting a significant practice effect.

Another concern is that we were not explicit enough when explaining our procedure. A number of participants tried to "outsmart" us and listen for superficial clues in the recordings, such as whether a real or synthesized piano was used, which evidently led them astray as both human-composed and computer-composed excerpts were created using the same equipment. We may also want to increase the duration of the excerpts as eight-bar phrases may be too short for listeners to be able to realistically gauge authorship.

We might consider abandoning the candid approach and instead employ an experimental paradigm that relies on deception, such as was done in Levisohn and Pasquier's evaluation of *BeatBender* (2008). This would rid us of the complications that arose from participants trying to over-dissect the musical excerpts for clues, and allow us to test for a larger range of properties. A limitation of our study was that it only asked participants to rate whether the pieces were human- or computer-generated; what, one could wonder, does this tell us about how successful the system was at being creative? If we modify the design and add additional criteria (e.g. ratings of *naturality*, *enjoyableness*, and *complexity* as was done in the evaluation of *BeatBender*) that parti-

cipants could listen for, it might tell us something more detailed about the differences between human-generated and computer-generated music.

Looking beyond the dichotomy of subjective judgements verses formalized criteria, there are arguably five levels of validation for artistic metacreations: the academic forum (whether the paper describing the creative system gets accepted or not), controlled evaluation (experiment such as those described in this paper), and feedback from journalists and critics, peers (artist from that community), and audiences. No evaluation study to our knowledge has attempted to cover all five of these levels. In future studies, we may consider rectifying this by adopting a methodology which would encompass all of these dimension, enhancing the validity of and confidence in our conclusions.

Acknowledgements

Thanks to Arne Eigenfeldt and David Mesiha for taking the time to give us a demonstration of the software and for providing troubleshooting correspondence.

References

Alwin, D. F., and J. A. Krosnick. 1991. *The reliability of survey attitude measurement:The influence of question and respondent attributes*. Sociological Methods & Research. 20 (1), 139–181.

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

Boden, M. 2010. *The Turing test and artistic creativity*, Kybernetes, 39(3), 409–413.

Brown, W. 1910. *The Judgment of Distance*. Publications in Psychology. 1, 1–71.

Cakewalk. http://www.cakewalk.com.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Creative Intelligent Systems: Papers from the AAAI Spring Symposium*, 2008. Menlo Park, CA: AAAI Press.14–20.

Drupal. http://drupal.org.

Eigenfeldt, A. and Pasquier, P. 2010. *Realtime Generation of Harmonic Progressions Using Controlled Markov Selection*. In *Proceedings of the First International Conference on Computational Creativity (ICCCX)*, ACM Press, Lisbon, Portugal, 16-25, 2010.

Eigenfeldt, A. and Pasquier, P. 2011. *Negotiated Content: Generative Soundscape Composition by Autonomous Musical Agents in Coming Together: Freesound*. In *Proceedings of the Second International Conference on Computational Creativity*, ACM Press, México City, México, 27-32, 2011.

Gilljam, M., and D. Granberg. 1993. Should we take "don't know" for an answer? *Public Opinion Quarterly*. 57, 348–357.

Harnad, S. 2000. Minds, Machines and Turing. *Journal of Logic, Language and Information.* 9(4), 425–445.

Hal Leonard Corp. 1993. *The Real Little Classical Fake Book*. Hal Leonard Publishing Corporation, Milwaukee, WI.

Levisohn, A. and Pasquier, P. 2008. *BeatBender: Subsumption Architecture for Rhythm Generation*. ACM International Conference on Advances in Computer Entertainment (ACE 2008), Yokohama, Japan, pages 51-58, ACM Press, 2008.

Native Instruments. http://www.native-instruments.com.

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

Pearce, M. and Wiggins, G. 2001. Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science*. Brighton: SSAISB. 22–32.

Pease, A., Winterstein, D., and Colton, S. 2001. Evaluating machine creativity. In *Workshop on Creative Systems, 4th International Conference on Case Based Reasoning*. (ICCBR-01), Vancouver, British Columbia, Canada, 30 July-2 August. 56–61.

Poe, G. S., I. Seeman, J. McLauglin, E. Mehl, and M. Dietz. 1988. "Don't know" boxes in factual questions in a mail questionnaire: Effects on level and quality of response. *Public Opinion Quarterly*. 52, 212–222.

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds & Machines*. 17, 67–99.

Sibelius. http://www.sibelius.com.

Spector, L. and Alpern, A. 1994. Criticism, Culture, and the Automatic Generation of Artworks. In *Proceedings of Twelfth National Conference on Artificial Intelligence* (Seattle, Washington, USA, 1994). 3–8. AAAI Press/MIT Press.

Stamp, A., Isenberg, T., and Carpendale, M.S.T. A Case Study from the Point of View of Aesthetics: A Dialogue Between an Artist and a Computer Scientist. In *Proceedings of Computational Aesthetics in Graphics, Visualization, and Imaging 2007* (CAe 2007, June 20–22, 2007, Banff, Alberta, Canada). (Aire-la-Ville, Switzerland), Eurographics Association. 129–134, 2007.

Turing, A. 1950. *Computing Machinery and Intelligence*. Mind. 59, 236 (Oct. 1950), 433–460.

Wallraven, C., Cunningham, D., Rigau, J., Feixas, M. and Sbert, M. 2009. Aesthetic appraisal of art - from eye movements to computers. *Computational Aesthetics 2009: Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging,* 137–144.

Whitelaw, M. 2004. *Metacreation: Art and Artificial Life*. Cambridge, MA: MIT Press.