Investigating Listener Bias Against Musical Metacreativity

Philippe Pasquier, Adam Burnett, Nicolas Gonzalez Thomas
School of Interactive Art + Technology
Simon Fraser University
philippe_pasquier@sfu.ca

James B. Maxwell Arne Eigenfeldt
School for the Contemporary Arts
Simon Fraser University
jbmaxwel@sfu.ca

Tom Loughin
Department of Statistics and Actuarial Science
Simon Fraser University
tloughin@sfu.ca

Abstract

We present an empirical study investigating the hypothesis that listeners hold a bias against computer-composed music. Presented in part as a replication study, the proposed methodology seeks to improve upon weaknesses found in previous studies of the subject. Across two study periods, with approximately 60 subjects each, we failed to find evidence of a significant bias against computer-composed music. We outline potential weaknesses in our design, and propose improvements for future studies.

Introduction: Computational Creativity

A subfield of Artificial Intelligence (AI) that has recently gained significant momentum is the exploration of *computational creativity*. Pasquier et al. define this as "the science of machines addressing creative tasks" (Pasquier et al., 2016). Musical Metacreation (MuMe) is a subfield of computational creativity that addresses specifically musical tasks, like improvisation, composition, and performance. MuMe systems can be classified along a continuum according to their relative levels of autonomy, from completely user-dependent computer music "tools," to autonomous generative systems (Pasquier et al., 2016).

As an extension of AI, computational creativity presents a new set of philosophical difficulties. In the case of general intelligence, reasoning processes can often be understood through the *a posteriori* analysis of solutions—i.e., a kind of reverse engineering of thought. This phenomenon was noted by Minsky, who pointed out that once we understand how something is done, we no longer regard it as particularly intelligent, but instead see it as a straightforward, mechanical process (Minsky, 1982). Creative products, on the other hand—and perhaps more specifically the products of artistic creativity that we address here—do not necessarily obviate the processes from which they arise. In creative "reasoning," singularly optimal solutions seldom exist, and detailed knowledge of the technical means involved in producing a given solution cannot always account for the appropriateness of the method chosen.

Evaluation of the products of human artistic creativity is fundamentally subjective. In the field of music, the quality of works is evaluated by the artists themselves, by their peers (fellow composers and musicians), audiences (inferred from concert attendance and album sales), the media (published reviews of critics and journalists), programming requests from ensembles and presenters, and in some cases by the peer-review committees organized by public funding bodies and arts councils. In the academic world, research is likewise evaluated by the author and her/his peers, by acceptance at academic conferences and the receipt of grants and scholarships, but also via methods that strive for greater scientific objectivity; i.e., the measurement of formalized input/output relationships and empirical studies which record the responses of participants.

Scientific experiments seek to infer the presence of the investigated phenomena by contrasting an experimental manipulation with a baseline control. For metacreations, this control is comprised of human-created artistic works. When dealing with music composition, however, developing the control faces two significant challenges: 1) The problem of selecting a representative work which will contrast against the computer-composed work, while mitigating the introduction of confounding variables, and 2) The interpretational problem of presenting the music to listeners in a way that accurately communicates its essential qualities, while again limiting the introduction of confounding variables.

A Bias Against Musical Metacreation?

Evidence of the bias against computational creativity is perhaps best illustrated anecdotally. When David Cope debuted "Emmy"—Experiments in Musical Intelligence (Cope, 1996)—before a live audience, her compositions were reportedly panned by a critic weeks before the actual concert (Blitstein, 2010). Emmy was not a human composer, but rather a computer program developed by Cope to emulate the styles of famous composers. Poor reviews were not the only obstruction Emmy faced: audiences often reacted with anger to her works, record companies declined to sign recording contracts, and musicians refused to perform her music (Cope, 2004).

Undeterred by this negative response—some of which he described as "racist" for its anti-machine hostility—Cope developed the software further (Blitstein, 2010). This work culminated in a successor to Emmy, called "Emily Howell," which turned away from style emulation, focusing instead on the creation of novel works in a unique musical style. Some reviewers complained that Emily's works, though musically pleasing, were hollow, shallow, and lacking depth and heart (Cope, 2004). Cope lamented the bias of his critics, and their complaints that he was taking away one of the remaining things humans could claim was uniquely their own: *creativity* (Cheng, 2009).

Defining "Bias"

For the purposes of the current investigation, it is important to be clear about our definition of "bias." Explicit biases are consciously held attitudes or preferences, that can be directly reported by individuals. Implicit biases—also referred to as "cognitive" biases (Tversky and Kahneman, 1974)—on the other hand, are unconsciously held beliefs or attitudes that are not directly accessible to subjects, and thus influence behaviours and choices without the individual's awareness. A typical example might be a CEO who verbally states (and may consciously *believe*) that applicant gender should not influence hiring strategy, yet whose hiring patterns show a strong gender preference.

Through this study we seek to experimentally determine whether there is, or is not, a bias against the notion of computational creativity in music. We do not address explicit/implicit bias directly, but rather attempt to determine whether a general bias exists, such that knowledge of authorship alone—human or computer—is enough to significantly modify preference.

Replication Studies

Reproducibility is an essential principle of scientific research. Statistical convention considers significant results to be those that are unlikely to be attributable to sampling error alone—e.g., the difference between the true population mean and the mean of the sample being tested. The significance value (p) is used to indicate whether the observed effect may be attributed to sampling error alone, such that "significant" results indicate that sampling error should account for the effect $< p * 100\%$ of the time. However, where the responses of human subjects are involved, explicit test methodology, experimental design, and subject selection comprise a complex web of influences and interactions, such that it isn't always clear that the effect being observed is, in fact, the effect under investigation. This phenomenon is called "confounding," and it plagues observational studies especially, but also experiments with inadequate controls. Thus, while statistical testing may suggest that an effect is unlikely to be due to sampling error alone, it cannot rule out alternative influences as potential causes. Replication studies help ensure that the results found in one study can be reproduced at a later date, thereby building confidence in the verity of the observed result.

An example of the importance of reproducibility occurred in 2011, when social psychologist Daryl Bem published a study demonstrating so-called "psi" abilities in the Journal of Personality and Social Psychology (Bem, 2011). In the study, participants appeared more likely to recall words from a word list if they practiced typing out those words *at a later date*—i.e., they seemed to demonstrate a kind of "premonition" of the *future* rehearsal process that could improve their recall scores in the present. Perhaps unsurprisingly, these results failed to be replicated by subsequent studies (Young, 2012), rekindling conversation about the importance of replicability in the sciences.

Aside from deliberate manipulation, fraud, and statistical chance—such as that due to type-I error-rate (i.e., "false-positive") inflation arising from multiple comparisons (Garcia-Marques and Azevedo, 1995)—there are a number of reasons for this decline in replication studies. Broadly speaking in the contemporary research milieu, a market-driven mentality has encouraged an over-emphasis on novel, exciting, and often counter-intuitive results, consequently discouraging both the submission and publication of articles with negative findings; a phenomenon known as the "file drawer effect" (Rosenthal, 1979). In some cases, of course, replication studies that fail to obtain the significant results of an original study do so as a result of methodological imprecision. Some argue, however, that such "conceptual" replications are preferable, as they interrogate the generalizability of the phenomenon being studied (Young, 2012)—i.e., by reframing the intent of the study independently of its precise methodology. Another possibility is whether there has simply been a change in the cultural zeitgeist; i.e., is difficulty replicating behaviour data from past decades attributable to a genuine decline in the studied behaviour, thus indicating a shift in attitudinal norms? For instance, we would not consider the data acquired from an early 20th century study of Western attitudes toward homosexuality to be representative of the population today, and the inferences one could draw from such data would be similarly inapplicable.

The Replicated Study: Moffatt and Kelly

Empirically, (Moffat and Kelly, 2006) studied the proposed bias against computational creativity in the context of computer-composed music. In this study, participants listened to six one-minute musical excerpts, half of them human-composed and the other half computer-composed. The pieces were in three different "styles": "free-form jazz", "strings", and "Bach." To minimize effects from participants' personal preferences, the pieces were presented in three pairs: one human- and one computer-composed for each style. The selection of pieces was determined by their "surface similarity"[1] in an effort to conceal their authorship (Moffat and Kelly, 2006).

A group of 20 participants were divided into "musician" and "non-musician" groups, based on their level of formal music training and/or experience. Participants listened to each of the compositions and indicated how much they "liked" the composition, on a 5-point Likert scale, and whether they thought it was human- or computer-composed. After this initial round of listening and evaluation, the origins of the pieces were revealed and the participants were asked to evaluate the pieces again. In the second round, the questions were disguised so as to not alert participants to the purpose of the study—in this case asking them how willing they would be to buy, download, or recommend the compositions to someone, and how much they "enjoyed" each piece.

The experimenters noted that participants appeared to demonstrate a prejudice against computer-composed music, generally preferring those pieces that they believed (by their own judgement) to be human-composed. The experimenters dubiously called this the "free prejudice effect": participants were "prejudiced" in favour of pieces they freely decided were human-composed. However, the experimenters did not find any overt prejudice: there were no statistically significant drops in the evaluations of the computer-composed pieces upon revelation.

They also found that participants were able to identify the

[1] Details regarding their definition of surface similarity—i.e., the particular musical features considered—are not provided.

computer-composed pieces as computer-composed, and that non-musicians out-performed musicians at this task. Non-musicians, however, were not statistically successful at identifying human-composed music as human-composed. Participants, both musician and non-musician, also preferred human-composed pieces over computer-composed pieces, regardless of what they guessed their authorship to be, with musicians preferring the human-composed pieces to a greater degree. It is worth noting, however, that Moffat and Kelly draw several unsubstantiated and ill-informed conclusions from their data; an ethical problem we seek to avoid in the present study.

Burnett, Khor, Pasquier, and Eigenfeldt

In a previous study, Burnett et al. (Burnett et al., 2012) addressed similar questions in their evaluation of a system for generating harmonic progressions (Eigenfeldt and Pasquier, 2010). This experiment had certain methodological aspects in common with Moffat and Kelly, dividing participants into musician and non-musician groups, and using a Turing Test-like paradigm to determine whether participants could discriminate between human- and computer-composed musical excerpts. Whereas the Moffatt and Kelly study used deception—concealing the purpose of the study from participants—Burnett et al. explicitly informed participants that they would be listening to a mix of human- and computer-generated harmonic progressions and that they would be asked to identify the source of each (thus mirroring the general objective of the original Turing Test). Additionally, Burnett et al. sought to estimate the confidence of participant responses through the use of a 4-point Likert scale: 1) Definitely Human, 2) Probably Human, 3) Probably Computer, and 4) Definitely Computer.

The findings of Burnett et al. echoed those of the earlier study in a variety of ways. For example, like Moffat and Kelly, Burnett et al. discovered that non-musicians outperformed musicians at discriminating between the human- and computer-composed excerpts. However, whereas Moffatt and Kelly found that participants more easily identified the origin of the computer-composed works, Burnett et al. found the opposite; participants struggled to identify computer-composed pieces as computer-composed, but generally succeeded at identifying human-composed pieces. With regard to the measure of participant confidence, no significant differences were found between musicians and non-musicians, but participants were generally more confident in their evaluations of pieces that were human-composed. However, it is difficult to make direct comparisons between the studies, given these divergent results, as aesthetic and stylistic differences between the musical materials used in each study throw into question the influence of authorship on listener evaluations.

It is worth noting that both Moffat and Kelly and Burnett et al. received feedback indicating that participants attempted to "outsmart" the experimenters, listening for clues that would reveal the true authorship of the excerpts. In both cases, it was assumed that this effort mislead them into giving incorrect responses.

Experimental Methodology

With the possible exception of (Moffat and Kelly, 2006), there is a lack experimental data corroborating the presence of a bias against computational creativity in music. Here we describe an experimental attempt to determine whether such a bias exists. This experiment is, in part, a replication study, but it also attempts to improve upon previous studies, taking into account the deficiencies of both Burnett et al. and Moffatt and Kelly.

As indicated by participant comments, these previous studies encountered difficulties with participants trying to outsmart the experimenters by listening for "clues" of authorship—a problem we attempt to reconcile with this new procedure. Douglas Hofstadter noted that the potential for a bias against machine creativity might make it necessary to purposely deceive listeners as to the origins of a piece of music (Cope, 2004). The employment of deception in the current experiment has been designed to determine whether this speculation was correct. Efforts have also been made to reduce any practice effects stemming from the non-randomized presentation of the musical excerpts; a problem that affected previous studies. Familiarity and listening fatigue effects were also minimized by including a control condition, which allowed us to track changes in participant evaluations in the absence of any experimental manipulation.

To address problems of music selection, we attempted to reduce the perceptual differences between musical pieces by limiting them to a single instrumental timbre: contemporary string quartet. The three computer-composed samples were excerpts from two longer works and exhibited three distinct musical textures: homophony, polyphony, and heterophony (the simultaneous variation of a single melodic line). The generative systems for these works are described in (Eigenfeldt, 2012) and (Eigenfeldt, Burnett, and Pasquier, 2012). Although both generative systems were corpus-based in some way, many of the musical decisions (i.e. voice-leading) were based upon an auto-ethnographic analysis. Eigenfeldt selected two excerpts from his own music that matched the textures and harmonic language in two of the generative works, and composed a new excerpt to match the missing one. We also paired computer- and human-composed pieces based on shared structural aspects, taking into account tempo, rhythm, and dynamics. We believe this approach offers a significant improvement upon Moffat and Kelly's notion of pairing works by so-called "style", particularly given the rather unlikely pairings they chose.

To address the "interpretational problem", all pieces were performed by live musicians, in an effort to normalize the musical percepts and reduce variability. Further, in recording the six excerpts, the musicians did not know which works were human-composed or computer-composed, and each excerpt was allocated an equal 30 minutes for rehearsal and recording.

Methodology

Participants

Participants were recruited from SFU using dissemination emails that were sent out to the School for the Contemporary Arts, SIAT, Cognitive Science, and Psychology programs. Participants were incentivized by informing them that they would be included in a draw for four $50 cash prizes upon completion of the study.

Unlike previous experiments which tested participants' ability to discern the origin of a composition in a Turing-

like test, we were now interested in whether participant beliefs about the origins of the pieces had an effect on their evaluations. It was therefore no longer necessary to identify musicians and non-musicians within the pool of participants. However, we believed (as did Moffatt and Kelly) that a participant's cultural, academic, and musical background might help elucidate the reasons for the extent (or presence) of any bias they might have. Therefore, part of the experiment included a demographic questionnaire requesting that each participant indicate their age, gender, university major, country of birth, number of years residing in Canada, number of years studying/playing music, and number of years experience with computer programming languages (as an indication of their computer literacy).

The experiment was run on two separate occasions (Studies 1 and 2 below), with 60 subjects participating in Study 1, and 62 subjects participating in Study 2. Both studies utilized the same experimental design and online test interface, and were methodologically identical.

Presentation of Musical Examples

The programme of musical works was derived from video recordings of live musicians preforming three computer-composed and three human-composed musical works. The pieces presented were composed by, or generated by software designed by, Arne Eigenfeldt. All pieces were performed by the Yaletown String Quartet in Vancouver.

Participants viewed six video recordings, approximately one minute in length each, of the quartet performing each of the pieces. This method of presentation was used, in part, to address a deficiency in previous experiments. Granting the participants the ability to see human musicians performing the compositions was intended to help eliminate some of the listening "strategies" participants had previously employed to determine composition authorship—e.g., believing that subtle variations in the quality of the audio betrayed the composition's origin. Having all pieces performed by human musicians not only "normalizes" the quality of the recordings, it also presents the excerpts in a more realistic setting, potentially allowing us to more accurately capture participants' perceptions of the music.

Procedure

Participants were provided with a URL to an online survey. The survey was built using Drupal (drupal.org), with additional modules to enable audio and video playback and time tracking (i.e., to ensure that the participants listened to the musical excerpts in full). Participants were then presented with a consent page indicating that completion of the survey would constitute consent.

Participants were presented with one piece of music at a time. After each musical excerpt, they listened to a 10-second "palette cleansing" recording of the musicians tuning their instruments, to help reduce context effects that would arise from directly following one piece with the next. Practice effect was minimized by randomizing the order of presentation.

After each video presentation, participants were asked to indicate their impressions of each piece, on four different attributes, by ranking them on a bipolar scale with 50 discrete points, labelled only at the left and right extremes. The

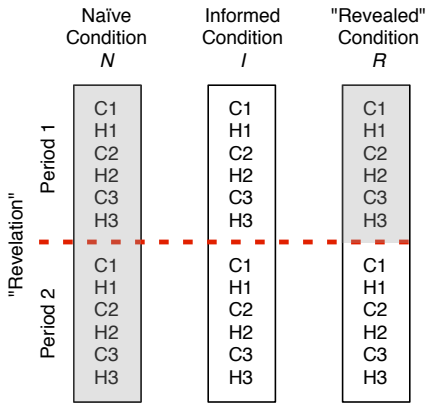

Figure 1: Experiment design using the 3 conditions. Grey shading indicates listening periods during which subjects are unaware of authorship. The horizontal line represents the "revelation" moment, where the second round of evaluations begins. "H" and "C" indicate (un-randomized) human- and computer-composed pieces, respectively.

following dimensions were indicated: **Good-Bad**, **Like-Dislike**, **Emotional-Unemotional**, and **Natural-Artificial**. This evaluation procedure was inspired by that used to evaluate the BeatBender metacreation (Levisohn and Pasquier, 2008). The spreading out of evaluations in this manner (i.e., using a set of bi-polar pairs) was proposed to help identify bias manifesting in a way that could have been obscured were the evaluations condensed into a single rating variable such as "liking" (as in previous experiments). Additionally, past research has indicated that maintaining focused attention (along with discriminative listening and emotional involvement) is critical for the accurate assessment of musical aesthetics (Madsen and Geringer, 2008). In marketing research, predictive validity has been shown to be high when using multiple-item scales over single-item scales (Diamantopoulos et al., 2012). We believe having participants contemplate this variety of dimensions during the listening task is an appropriate way to facilitate the desired level of attention and involvement.

The full set of musical excerpts was presented to participants twice, each time under one of two experimental settings. In the "naïve" setting, participants were not informed of the authorship of any of the pieces, and the experiment was presented as an investigation of the effect that visually witnessing a performance has on one's aesthetic evaluation of the music performed. In the "informed" setting, however, participants were explicitly told (and reminded) of the authorship of the pieces.

Ideally, the two exposures to the excerpts would have taken place under all possible combinations of settings to eliminate confounding covariates. However, we cannot deinform participants about piece authorship once they have been informed. Therefore there were three different experimental conditions: fully naïve (N), fully informed (I), and "revealed" (R). For the "revealed" condition, subjects start the experiment in a naïve condition, but conclude in an informed condition, with a revelation occurring midway (see Figure 1). This allows us to check for any "reaction" effect, where the shock of the revelation inspires a drastic change in evaluations.

Having these three groups, each broken into two periods, wherein the six pieces are reevaluated, also allows us to control for novelty, exposure, and fatigue effects. If we had conducted the experiment solely with the "revealed" group, as in Moffatt and Kelly, any increase or decrease in evaluations could be attributed to a loss of novelty (i.e., becoming bored), or an increase in familiarity (the "mere exposure" effect, leading to an increase in enjoyment). Listening fatigue could affect the second evaluation in unpredictable ways.

Following the second round of evaluations, participants were thanked for their participation but were asked one final demographic question about their experience with computer programming languages. This was asked at the end of the experience so as to not rouse suspicions and tip participants to the true nature of the study during the initial demographic questions. Participants were then directed to a separate website where they provided their email address so that we could confirm they had finished the survey. Here they were given the option to indicate whether they would like to be contacted about the results of the experience and/or be entered into the prize draw. As this section was separate from the survey-proper, it prevented us from matching survey answers to identifiable e-mail addresses, preserving anonymity.

Despite the differences between our two designs, we believe the design of the present experiment is similar enough to that used in (Moffat and Kelly, 2006) to allow for easy comparisons: both contain three human and three computer-composed pieces, in our case paired by composer rather than style (all six pieces in the present experiment were composed for string quartet). Our mixed (R) condition mimics that used in Moffatt and Kelly, but our addition of the fully naïve and fully informed conditions allowed us to check for timing and fatigue effects as well. Our experiment also had the advantage of asking the same evaluation questions in both rounds: Moffatt and Kelly asked participants how much they "liked" the compositions in the first round and how much they "enjoyed" the pieces in the second. However, the interpretation of these words is entirely subjective and could vary widely across participants (one can "like" a song very much, but depending on their current mood, they may not actually "enjoy" it at that particular moment). This was a concern that Moffatt and Kelly themselves expressed. We attempted to devise our cover story such that we could ask the same questions during both rounds without the similarities cuing the participants to the true nature of our experiment and compromising the validity of the results. The Moffatt and Kelly study also suffered from a lack of randomization of excerpt presentation order, and this too has been addressed for the present study. Finally, the addition of the demographic questionnaire inquires about the age, sex, and culture of the participants, factors which Moffatt and Kelly believed could shed light on the proposed bias we are seeking to identify.

Hypotheses

We anticipated a number of effects, both within and between the three conditions N, I, and R. The null hypothesis is that we should see no significant differences in the evaluation of the pieces among the three groups; the only changes between the initial and subsequent presentations of the stimuli should reflect novelty, exposure, or fatigue effects and have the same influence in each condition. As for anticipated experimental effects, we formed four main experimental questions:

1) Among those who hear the pieces naively in the first period, does the comparison of human- vs. computer-composed music change their ratings by different amounts when the authorship is revealed before the second hearing than when it is not?

2) In the second hearing only, is the comparison of human- vs. computer-composed music different when the authorship is known than when it is not?

3) Among those who hear the pieces naively in the first period, is the comparison of human- vs. computer-composed music different when the authorship is known than when it is not?

4) Does the comparison of human- vs. computer-composed music change by different amounts when the authorship is revealed partway through than when conditions remain fixed for both hearings?

Statistical Methods

Experimental Design

We created two sets of example pieces; the Human set (a, c, e) and the Computer set (b, d, f). These pieces were paired, so that each human piece had a corresponding computer piece: $(a, b)(c, d)(e, f)$. Pieces (a, b) were denoted as `Pair` 1, pieces (c, d) as `Pair` 2, and pieces (e, f) as `Pair` 3. These three pairs were always presented in this order, but within each pair the order of human or computer composition was considered both ways, resulting in 8 different versions of the survey.

Subjects were assigned to one of these 8 sequences upon enrolment into the study. The experiment was therefore conducted as a split-split-split plot crossover design. `Group` (N, R, or I)—a fixed effect—was assigned to a subject—a random effect. Within each subject there were 12 hearings arranged in nested groups of decreasing size. The fixed effect `Period` has two levels representing the six hearings before the potential reveal (`Period` $= 1$) or after (`Period` $= 2$). Within each `Period`, the factor `Pair`, with levels 1, 2, and 3, is assigned to the two hearings representing the pairs of compositions described above. We treat `Pair` as a fixed effect because the order in which the pairs of pieces are heard is always the same. Finally, within each pair, the fixed effect `Composer`, with levels H or C, is assigned to a hearing. Note that the key experimental factors, `Group` and `Composer`, are both randomized in this design. The fact that `Pair` was not randomized, but rather presented in sequence for each subject, (1, 2, 3) is unimportant because it represents a combination of the fatigue and other effects due to ordering of hearings and random variation among individual compositions. There is no interest in testing any part of this effect. Importantly, it does not confound with condition or composer.

Statistical Analysis

We analyzed the experiment using mixed-effect linear models according to its design (Milliken and Johnson, 2009), using JMP Version 12 software (2015, SAS Institute). We tested for carryover effects of this factor and found no significance.

The central questions our replication seeks to address are:

1. Whether people enjoy human music implicitly: H vs C in `Group` N
2. Whether people prefer human music when told: H vs C in `Group` I
3. Whether the difference of the above reveals a human- vs computer-music bias.

We organized our analyses into four contrasts, each designed to address a specific aspect of these hypotheses as described below. We applied the same contrasts to each response dimension.

Let H represent the model-estimated mean of a given response dimension for human-composed compositions, and C the mean of the same response dimension for computer-composed compositions. We use subscripts "1" or "2" to restrict these means to `Period` 1 or 2, respectively, and "N," "I," or "R" to restrict these means to `Group` N, I, or R, respectively. We use "Δ" notation to represent the change in mean responses before vs. after the potential reveal:

$$^{\Delta}H = (H_1 - H_2) \quad (1)$$

$$^{\Delta}C = (C_1 - C_2) \quad (2)$$

We add subscripts to these quantities to refer to these changes under a specific `Group`. All contrasts are tested against a t distribution with 714 degrees of freedom.

Results

In analyzing the data, we operated under the assumption that if results in the naïve condition showed no significant preference for either human or computer-composed music, then discrepancies in the other conditions may indicate the presence of bias. Note, however, that our focus is on the *response to being informed of authorship, not on the estimation of authorship itself.*

Broadly speaking, we considered three main factors: 1) The "pure musical impression" (represented by `Group` N), 2) The influence of knowledge of authorship (`Groups` N and R), and 3) The influence of the "reveal" (which takes into account an awareness of the deception).

Four contrasts were designed (series S_1 to S_4), based on the stated hypotheses above:

Series 1 isolates the difference across periods for the naïve (N) and mixed (R) groups:

$$S_1 = [^{\Delta}H_N - {}^{\Delta}C_N] - [^{\Delta}H_R - {}^{\Delta}C_R] \quad (3)$$

Since the individual terms represent changes of opinion (Δ), and each bracketed difference isolates the effect of authorship on that change, significant S_1 values could be said to indicate a bias—i.e., when informed of authorship, subjects change their opinions.

Series 2 compares only the second `Period` differences of I and R, to N:

$$S_2 = \frac{(H_2 - C_2)_R + (H_2 - C_2)_I}{2} - (H_2 - C_2)_N \quad (4)$$

Here we attempt to account for repetition effects, by evaluating only the opinions of subjects who have already heard the pieces. The evaluation is again between fully-informed subjects (i.e., since `Group` R in `Period` 2 is also informed) and naïve subjects, and compares the average evaluations of all informed subjects against those of the naïve subjects.

Series 3 is similar to series 2, but looking only at R vs N (i.e., excluding I), and thereby contrasting the purely subjective, musical impression with the knowledge of authorship:

$$S_3 = (H_2 - C_2)_R - (H_2 - C_2)_N \quad (5)$$

Series 4 looks again at difference across periods, contrasting the "control" `Groups` N and I, against R:

$$S_4 = \frac{[^{\Delta}H_N - {}^{\Delta}C_N] + [^{\Delta}H_I - {}^{\Delta}C_I]}{2} - [^{\Delta}H_R - {}^{\Delta}C_R] \quad (6)$$

Here, we could be said to most directly isolate the deception itself, since we normalize the change across periods for the pure musical impression (N) and the fully-informed evaluation (I), in the absence of any form of deception. These normalized "deception free" evaluations are contrasted with the R case, in which subjects transition not only from the pure musical impression to the knowledge of authorship, but *also* to an awareness of the deception (i.e., in `Period` 2, they become aware that half of their evaluations have been given with incomplete information). The combined results for both studies are given in Table 1.

	S_1	S_2	S_3	S_4
		"Emotional"		
t-ratio	-0.57	-0.27	0.63	-0.00
p-value	0.57	0.79	0.53	1.00
Scaled				
Estimate	-1.5	-0.4	1.2	-0.0
Std Error	2.6	1.5	1.9	2.4
		"Good"		
t-ratio	0.03	-0.52	-0.29	0.33
p-value	0.98	0.61	0.77	0.74
Scaled				
Estimate	0.1	-0.7	-0.4	0.7
Std Error	2.2	1.3	1.6	2.0
		"Like"		
t-ratio	-0.28	-0.52	-0.25	0.09
p-value	0.78	0.60	0.80	0.93
Scaled				
Estimate	-0.8	-0.8	-0.5	0.2
Std Error	2.7	1.6	1.9	2.5
		"Natural"		
t-ratio	0.47	-0.17	0.51	1.21
p-value	0.64	0.86	0.61	0.22
Scaled				
Estimate	1.2	-0.2	0.9	2.8
Std Error	2.5	1.5	1.8	2.3

Table 1: Summary of t-ratios, p-values, estimated contrasts, and standard errors for each series, across studies 1 and 2.

We ran a Factorial ANOVA with repeated measures and found no significance, affected primarily by small differences in means (<5 points) relative to scale size (50) and high variability. Additionally, we ran all pairwise comparisons of means for the `Group*Period*Composer` fixed effect, looking for patterns of possible mean differences us-

"Like"		
`Group, Period, Composer`	Least Sq Mean	Std Error
N, 1, C	19.5	1.3
N, 1, C	19.5	1.3
N, 1, H	20.0	1.3
N, 2, C	18.7	1.3
N, 2, H	19.5	1.3
R, 1, C	16.9	1.6
R, 1, H	17.6	1.6
R, 2, C	17.6	1.6
R, 2, H	18.0	1.6
I, 1, C	19.7	1.4
I, 1, H	21.0	1.4
I, 2, C	22.3	1.4
I, 2, H	22.0	1.4

Table 2: The Least Square Mean and Standard Error for the "Like" evaluation.

ing Tukey's Honestly Significant Difference (HSD) method, which controls the type-I error-rate inflation that can occur when multiple hypothesis tests are performed. There were no significant differences among means, indicating that none of the effects of these three factors, or their interactions, were particularly important to this study. The Least Square Mean and Standard Error results for the "Like" comparison are given in Table 2.

Discussion

As with the Moffat and Kelly study, the current study shows that the so-called bias against computational creativity, while observable, is mostly anecdotal and exaggerated While our results do indicate a negative effect of the knowledge of computer authorship on listener judgements, this effect is not significant.

A clear contributing factor in the failure to find significance is the small differences in means of the evaluations (<5 points on a 50-point scale), and the high variability within that range (see Figure 2 of the Appendix). This would appear to indicate a degree of uncertainty or ambivalence in the listeners, with regard to either the musical content of the examples, or the application of the given criteria to evaluating those examples—participants simply did not have strong opinions, regardless of their knowledge of authorship. Thus, although we did see the anticipated decrease in ratings on the "good" and "like" dimensions for the R group (i.e., after the "reveal"), the impact of this change was lost amid the general noisiness of the data.

Unlike Moffat and Kelly, we do not see a significant preference for human-composed music in the naïve group, N. We also found no significant decrease in preference for computer-composed music, among fully-informed subjects (group I). We did, however, see a slight skewing of scores toward the "bad" dimension in group R, after the "reveal", though the change was not significant. A similarly anticipated change toward "artificial" was noted along the "natural" dimension, though in this case the nature of the evaluation being made is a possible factor. Specifically, it is worth noting that this pair of labels, when compared to the others, fails to denote a clear positive/negative opposition—i.e., "bad" is clearly opposed to "good", "dislike" opposed to "like", and "unemotional" to "emotional." However, it is not so clear that "artificial" should be opposed to "natural." Though often used in opposition colloquially, these terms carry strong ideological associations (Žižek, 2011), and thus represent points along an ethical continuum, perhaps more so than extrema in a relationship of contradiction or reversal. The change in response in this case may be an indication of a cultural attitude toward these underlying ideologies—simply put, a human is "natural" and a computer "artificial"; the musical appraisal may be strictly coincidental.

Despite the lack of significant findings, it is worth making a couple of further observations, with reference to Figure 3 in the Appendix. First, there are notable differences between participant responses across the two studies. Looking at the dimensions related most directly subjective preference—"Good" and "Like"—we see that, while in Study 1 the results indicate a positive change in `Period` 2 for the N group, supporting empirical findings on the relationship between familiarity and musical preference (Szpunar, Schellenberg, and Pliner, 2004; Schubert, 2007), Study 2 opposes this pattern. In fact, Study 2 shows (weakly) correlated negative changes of rating in `Period` 2 across all groups for these dimensions, suggesting that perhaps listener fatigue is a contributing factor.

It is also worth noting that, in Study 1, for the "Natural" dimension, participants in the R group show a positive change for the computer-composed works, and a negative change for the human-composed works. This appears to suggest that they were pleasantly surprised at the "naturalness" of the works that were revealed to be computer-composed, and perhaps somewhat disappointed at the "unnaturalness" of the works that were revealed to be human-composed; an outcome that may point to the ideological underpinnings of the terms "natural" and "artificial" as a contributing factor, as discussed above.

Future Work

Although the sample sizes for the current study were statistically adequate, the narrow range of ratings, relatively high variability within that range, and overall change in ratings between studies 1 and 2, suggest that perhaps the study should be re-run with a larger number of subjects. Given the randomization of group assignment (N, I, and R), a larger overall population would also help balance the sizes of the different groups. We also note that use of the terms "natural" and "artificial" should be reconsidered, so as to indicate a clearer positive/negative opposition, in keeping with the other dimensions (i.e., "good-bad," "like-dislike," and "emotional-unemotional"). Additionally, we recognize that the standardizing choices made in the current study also represent a limitation, in that we have studied only these selected musical conditions. Hence, it cannot be confirmed that the results of this work would apply to other musical genres or other circumstances that differ from the conditions used here.

Further, we are interested in exploring the possibility of using rank-based, as opposed to ratings-based, evaluations. As outlined by Yannakakis and Martínez (Yannakakis and Martínez, 2015), rank-based questionnaires can help eliminate some of the problems associated with ratings-based questionnaires when evaluating subjective, psychological

factors like emotional response, preference, or opinion. In the case of the current study, for example, a rank-based choice—simply ranking the works in each `Pair` according to how well they represent the dimension under consideration ("good", "like", "emotional", "natural")—would eliminate the above statistical problem of narrowly distributed ratings, while also obviating use the word "artificial."

Finally, we are curious whether there may be a correlation between the rating-change across `Periods` (i.e., Eq. 1 and 2 above) and the musical content of the works under evaluation. Specifically, we are interested in knowing whether the overall complexity of the musical examples has an influence on listeners' ratings after the "reveal." We suspect that the degree of change may be correlated with the so-called "inverted-U" pattern, proposed to govern subjective judgements of aesthetic value (Walker, 1981). The inverted-U model suggests that the subjective assignment of aesthetic value follows an "inverted-U" pattern, such that value (or quality) is considered highest for works that match the subject's preferred complexity level—less complex works are rated lower, as are more complex works. We would like to know whether subjects that potentially hold a bias against computer-composed music provide more sharply contrasting evaluations after the "reveal" for works that match their preferred complexity level. This would suggest that it is not simply the fact of computers composing music that such listeners take exception to, but rather that such systems compose works with a complexity rivalling that of their preferred human-composed works. Such a finding would help establish a more adequately complex understanding of listeners' attitudes about computational creativity, while at the same time potentially offering a benchmark for selecting works—both human- and computer-composed—for similar studies in the future.

Conclusion

We outlined an experimental design for investigating the proposed bias against musical metacreativity. The experiment was conducted over two studies, on approximately 120 students from Simon Fraser University. Similar to a previous study by (Moffat and Kelly, 2006), we did not find significant support for the presence of such a bias, though our results do suggest that this bias exists, in some listeners. As our results were not significant, we refrained from conjecture based on demographic information. We also outlined a number of possible improvements to our design for future studies.

References

Bem, D. J. 2011. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology* 100(3):407.

Blitstein, R. 2010. Triumph of the cyborg composer. *Miller-McCune Magazine: http://www.miller-mccune.com/culture-society/triumph-of-the-cyborg-composer* 8507.

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

Cheng, J. 2009. Virtual composer makes beautiful music—and stirs controversy. *Ars Technica* 29.

Cope, D. 1996. *Experiments in musical intelligence*, volume 12. AR editions Madison, WI.

Cope, D. 2004. *Virtual music: computer synthesis of musical style*. MIT press.

Diamantopoulos, A.; Sarstedt, M.; Fuchs, C.; Wilczynski, P.; and Kaiser, S. 2012. Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. *Journal of the Academy of Marketing Science* 40(3):434–449.

Eigenfeldt, A.; Burnett, A.; and Pasquier, P. 2012. Evaluating musical metacreation in a live performance context. In *Proceedings of the Third International Conference on Computational Creativity*, 140–144. Citeseer.

Eigenfeldt, A. 2012. Corpus-based recombinant composition using a genetic algorithm. *Soft Computing* 16(12):2049–2056.

Garcia-Marques, T., and Azevedo, M. 1995. Multiple comparisons and the problem of alpha inflation. anova as an example. *Psicologia X* 195–220.

Madsen, C. K., and Geringer, J. M. 2008. Reflections on puccini's la bohème investigating a model for listening. *Journal of Research in Music Education* 56(1):33–42.

Milliken, G. A., and Johnson, D. E. 2009. *Analysis of messy data volume 1: designed experiments*, volume 1. CRC Press.

Minsky, M. L. 1982. Why people think computers can't. *AI Magazine* 3(4).

Moffat, D., and Kelly, M. 2006. An investigation into people's bias against computational creativity in music composition. In *Proceedings of the third joint workshop on Computational Creativity (as part of ECAI 2006), Rivia del Garda, Italy.*

Pasquier, P.; Eigenfeldt, E.; Brown, O.; and Dubnov, S. 2016. An introduction to musical metacreation. *ACM Computers in Entertainment*.

Rosenthal, R. 1979. The file drawer problem and tolerance for null results. *Psychological bulletin* 86(3):638.

Schubert, E. 2007. The influence of emotion, locus of emotion and familiarity upon preference in music. *Psychology of Music* 35(3):499–515.

Szpunar, K. K.; Schellenberg, E. G.; and Pliner, P. 2004. Liking and memory for musical stimuli as a function of exposure. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30(2):370.

Tversky, A., and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science.* 185(4157):1124–1131.

Walker, E. L. 1981. The quest for the inverted U. In *Advances in intrinsic motivation and aesthetics*. Springer. 39–70.

Yannakakis, G. N., and Martínez, H. P. 2015. Ratings are overrated! *Frontiers in ICT* 2:13.

Young, E. 2012. Nobel laureate challenges psychologists to clean up their act. *Nature News*.

Žižek, S. 2011. *Living in the end times*. Verso.

Appendix

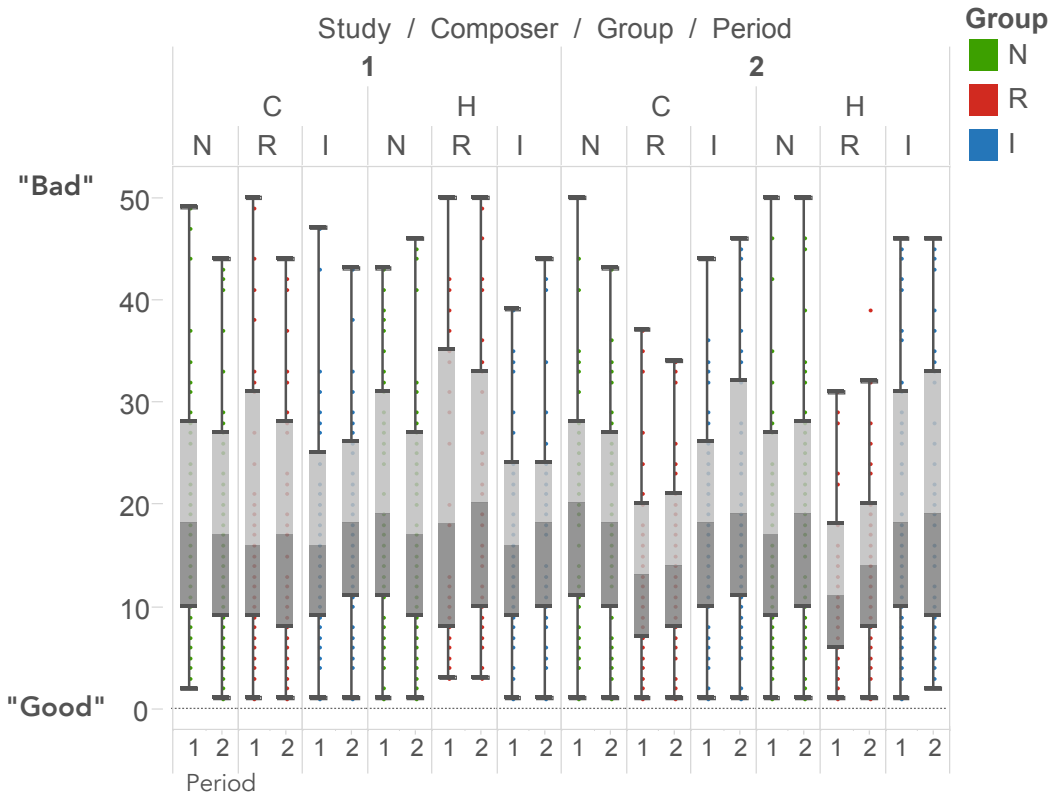

Figure 2: Bar graph of subject ratings on the ”Good” dimension.

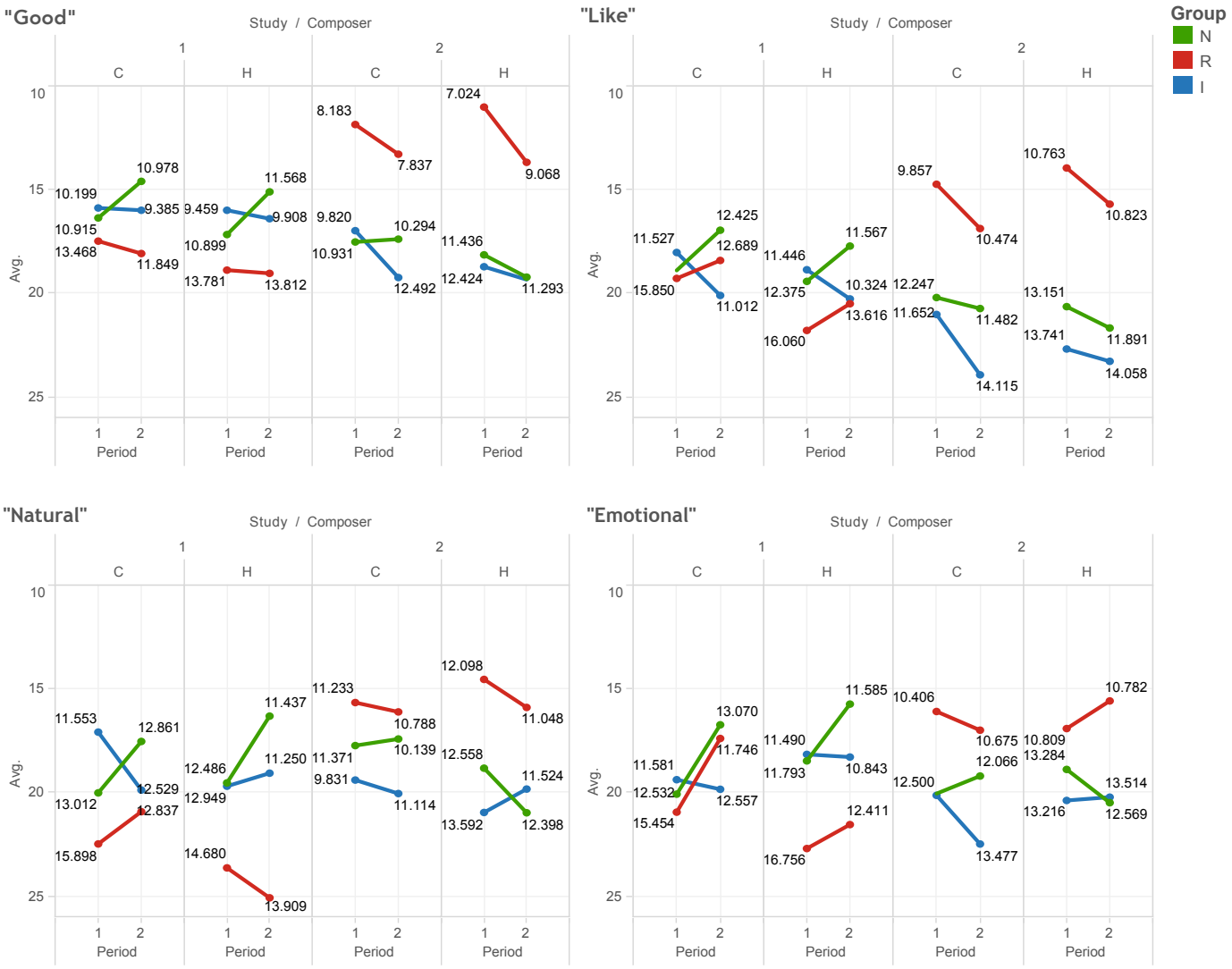

Figure 3: Mean subject ratings for all dimensions. Note that the y-axis is scaled to focus on the range of subject responses. The actual scale range for the study was 0 to 50, with 0 on the left and 50 on the right—i.e., 0=left="good", 50=right="bad".