A. Georgaki and G. Kouroupetroglou (Eds.), Proceedings ICMC|SMC|2014, 14-20 September 2014, Athens, Greece

Music Systemisers and Music Empathisers – Do they rate expressiveness of computer generated performances the same?

Emery Schubert
UNSW Australia
e.schubert@unsw.edu.au

Giovanni De Poli
University of Padova, Italy
depoli@dei.unipd.it

Antonio Roda
University of Padova, Italy
roda@dei.unipd.it

Sergio Canazza
University of Padova, Italy
canazza@dei.unipd.it

ABSTRACT

This paper investigates three individual differences with respect to ratings of the same piece of classical piano music that has undergone different expressive performance treatments. The individual difference variables investigated were music systemising (those interested in the structural and organizational aspects of music), music empathizing (those interested in the emotional/human aspects of music) and musical experience (years of playing). Five pieces, based on stimuli used in Rencon-GATM were rated according to expressiveness and execution, each being related to musical expression, but the former suggesting an empathizing processing style and the latter a systemizing processing style. Ratings made by 45 participants did not show any clear differences that could be attributed to a cognitive style. One explanation for this finding was that cognitive music styles are more likely to influence justifications of ratings, rather than ratings magnitude. High music systemisers reported having higher concentration than other participants.

1. INTRODUCTION

Expressiveness is a critical factor in determining whether one performance of a piece is better or more enjoyable or more interesting that another performance of the same piece. The performer typically manipulates a number of musical parameters (such as timing and dynamics) to achieve expressive nuances. However, performance rules have been identified which are thought to correlate with appropriate levels of expression for a given style of music [1]. In recent years, these rules with or without human intervention have enabled programmable, computer generated performances to sound more and more convincing to listeners as authentic and expressive [2-4].

However, as these algorithmic performances become more sophisticated, the question of audience response must arise. Even with traditional performances of the Western canon individuals differ in their judgments of the same performance. We wanted to explore whether individual differences influence judgements about different models of expressiveness generated or assisted by algorithms which control timing and velocity of keystrokes on a piano, for example via a diskclavier. Recent renewed interest in cognitive styles [5], and in particular in those specifically related to music were thought useful as a starting point. Cognitive style measures as applied to music are based on earlier, more general work by Baron-Cohen which is concerned with extreme male brain theory and autism [6-8]. Music cognitive styles [9] consist of two subscales: Music Empathising and Music Systemising. Music empathisers (ME) are characterized by an interest in the emotional/human aspects of music, and thus from a naïve perspective, one might postulate that such individuals do not exhibit a strong affinity with musical expression generated by a computer model [10]. In addition, they may prefer to focus attention towards expressive, emotional aspects of the performance. Music systemisers (MS) are interested in how music works, its structure, form and statistics. Again, from a simple, naïve perspective, such individuals would be interested in computer generated performances, and so may respond positively toward them. Furthermore, they should prefer to focus on technical aspects of a performance.

A third individual difference variable investigated was musical experience. Having music experience as a variable allowed two matters to be addressed. First, we could examine whether more musically trained people made more consistent responses than less musically trained people, and second we could check for similar trends between music cognitive styles and music experience to reduce the risk of making conclusions based on a confounding variable. Research by Kreutz et al [9], for example, suggests that musical experience is related to music systemizing to a greater extent than it is to music empathizing.

2. AIMS

This study aimed to explore whether high music systemising individuals and high music empathizing individuals rated expressiveness of different performances differently to their low cognitive style counterparts. In this study we restricted our investigation to the simple

rating of expressiveness in two way, one which might encourage music empathising responses (via rating of 'expressiveness' and another that might encourage music systemizing responses (via rating of the execution of the performance).

3. BACKGROUND TO THE STUDY

This paper reports the expressiveness ratings made by students in Sydney, Australia in 2013 using four sound recordings of Allegro Burlesco Op. 88 by Kuhlau produced for Rencon-SMC11 [11], plus an additional recording by a human performer made in Bologna. Due to space constraints, readers are invited to inspect Canazza et al [11] for background information about the Rencon-GATM project.

The aspect of the Sydney study reported here is part of a larger project investigating individual differences in judgements of computationally generated expressiveness models. The participants were requested to respond to a number of questions for each excerpt as per Rencon-SMC11. Furthermore there was no human performance at Rencon-GATM, just the four computer generated pieces.

A key aim of the Rencon-GATM project was to determine which realization of the Kuhlau was rated as the most effective from a musical expressiveness point of view. The present study continues examining more detailed aspects of individual response reported in the Bologna data set [11] to preference for different computer generated and human renditions. In that study, gender and music cognitive styles were examined, but indicated no significant differences between the two groups. One reason for lack of effect may have been due to the small variance in the music cognitive style variance. The participants in that study had rather high music systemizing scores overall, for example [9].

4. METHOD

The stimuli were presented in the sequence Perf1, Perf2, Perf3, Perf4, Perf5 (human), followed by the first four stimuli presented again in the same order (hence '1234h1234'). The first four pieces are referred to as 1a, 2a, 3a and 4a respectively. When played the second time they are referred to as 1b, 2b, 3b and 4b respectively. The systems used to generate the version are [11, p. 354]:

- 1a: uses two algorithms: YQX , developed by Dept. of Computational Perception, J. Kepler University, Linz (Austria) for tempo and Basis mixer for dynamics;
- 2a: CaRo 2.0, developed by the Sound and Music Computing group, Dept. of Information Engineering, University of Padova (Italy);
- 3a: DirectorMusices, developed by the Music Acoustics Group, KTH Royal Institute of Technology, Stockholm (Sweden);
- 4a: VirtualPhilharmony, developed by Katayose Lab., Dept. of Human and Systems Interaction

The human (h) performance is presented once only in the sequence. Participants completed the study via KeySurvey survey software (https://www.worldapp.com/surveys/overview.html), at their own pace on their own computer/sound-system. They were not told that some of the pieces were repeated.

Forty-five participants took part in this particular study in return for course credit – consisting of 31 females, 14 males, with overall mean age of 21.4 years (range 18-34), and overall average mean years of playing a musical instrument of 6.3 (range 0-16). The participants listened to each of the 9 stimuli and rated a number of qualities on a scale of 0 to 10 for each piece. A rating of 10 indicates a very strong agreement with the item, and a rating of 0 a complete disagreement with item. The study was conducted over the internet, and participants were asked to complete the study in a private, quiet space with good quality speakers or headphones. Participants were asked to report the audio output equipment they used. The qualities rated were: enjoyment of performance, enjoyment of piece, expressiveness, execution of performance, played by human, played by robot, familiarity with piece, with performer, task concentration, and equipment quality. For space reasons, results for only the most pertinent response qualities are reported here. Specifically, the results for two items are presented: 'The performance was expressive' and 'The execution of the performance was good (well played)'. Since nine, roughly two-minute pieces are rated, participant concentration could be a crucial variable, and so self-reported rating of concentration is reported (10 being high and 0 being low) for each stimulus. Music empathising and music systemizing ratings were collected in a separate survey sent to the same participants approximately two weeks earlier, administering the Music Cognitive Style scales [9].

5. DATA PREPARATION AND ANALYSIS

Analysis comparing groups used median split scores for music systemizing, music empathizing and years of playing instrument ('Musician'). The Above median group for each variable is referred to as the **A group**, and the below median group is referred to as the **B group**. The groups were determined post hoc. In all figures that follow, error bar pairs should be read as A group for the solid line on the left of the pair and B group for the dashed line on the right of the pair.

6. RESULTS AND DISCUSSION

6.1 Expressiveness

Overall, stimulus 2a (CaRo) was given the highest rating of the 9 stimuli with a mean of 7 to 7.5 (left pane), but performance 4a (VirtualPhilharmony) is rated quite erratically, with mean ranging from around 6 up to nearly 8 out of 10. There is unlikely to be any main effect as to the most expressive performance, but the human (h) performance had the highest mean rating overall.

A trend can be observed in cognitive music style, with A group ME scoring expressiveness higher than the B group ME. 4a was rated as having the mean highest expressiveness for the ME-A group and the Musician-A group. MS-A liked 2a the most. The B levels for each group gave overall lower ratings for expressiveness (that is, ratings closer to the neutral 5 position on the 0-10 scale). Although these results suggest an effect of music empathising, the same trend in results can be observed for MS, with the A group generally rating expressiveness higher than or the same as the B group, as well as the Musician group (A group rating expressiveness the same or higher than the B group). Therefore the high ratings by ME indicate that expressiveness ratings are not related to music empathising, or that they are mediated by other factors.

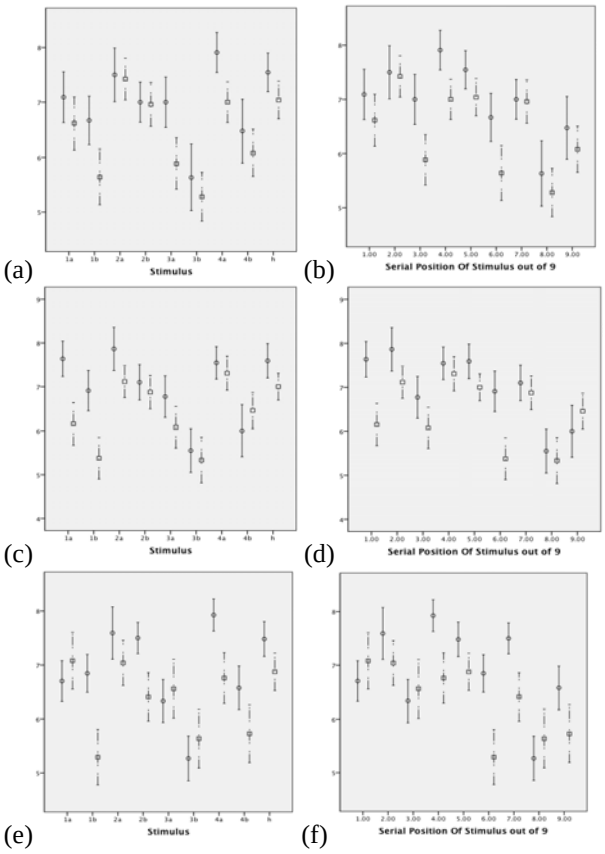

Figure 1. Error bar plots for expressiveness ratings by stimulus.
(a) ME by stimulus, (b) ME by serial order, (c) MS by stimulus, (d) MS by serial order, (e) Musician by stimulus, (f) Musician by serial order.
Error bar = ±1SE. Solid line is A (above median) group, Dashed line is B (below median) group.

6.2 Execution of performance

Technical executions of the pieces were generally high (all means ratings above 6/10, and 13 means were above 7). MS-A reported performance 3b (DirectorMusices) as being fairly low in quality of technical execution. While we might hypothesise that MS-A participants will be good at rating technical execution, the graphs demonstrate that the low rating of 3b is inconsistent with 3a, which is rated as much higher [by error bar inspection. See 12, 13], even though they were the same performance. The poor reliability is most likely generated by mental fatigue effects [14]. And so we inspected possible fatigue effects through analysis of self-rated task concentration.

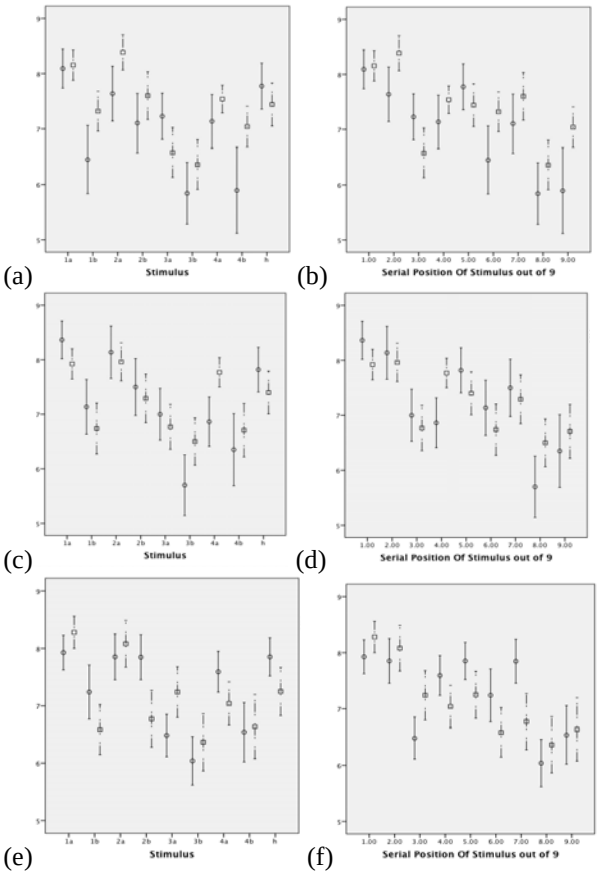

Figure 2. Error bar plots for performance execution ratings by stimulus.
(a) ME by stimulus, (b) ME by serial order, (c) MS by stimulus, (d) MS by serial order, (e) Musician by stimulus, (f) Musician by serial order.
Error bar = ±1SE. Solid line is A (above median) group, Dashed line is B (below median) group.

6.3 Concentration

ME-A group appears to be more consistent with concentration ratings, maintaining it at a higher level after the fifth (serial) performance compared to ME-B. MS-A reported the highest level of concentration overall, but particularly during the first four excerpts, relative to all other groupings.

7. GENERAL DISCUSSION AND CONCLUSION

Although some trends were observed, neither music systemising or music empathizing could be implicated in rating differences for either the expressiveness or the technical execution of the five performances investigated. For example, even though ME-A (high music empathiz-

ing participants') ratings of expressiveness of stimuli were overall the same or higher than ME-B, the same trend was observed for the MS groups and the Musician groups. Subsequently we proposed the following conclusions:

1. Music cognitive styles do not influence preference ratings because they are a reflection of a style of processing—that is the justification of the judgement, and not the magnitude of the judgement itself. Some preliminary evidence for this conclusion can be found in De Poli et al [10], although a recent study [15] suggests that justifications are not separable according to either music cognitive style.

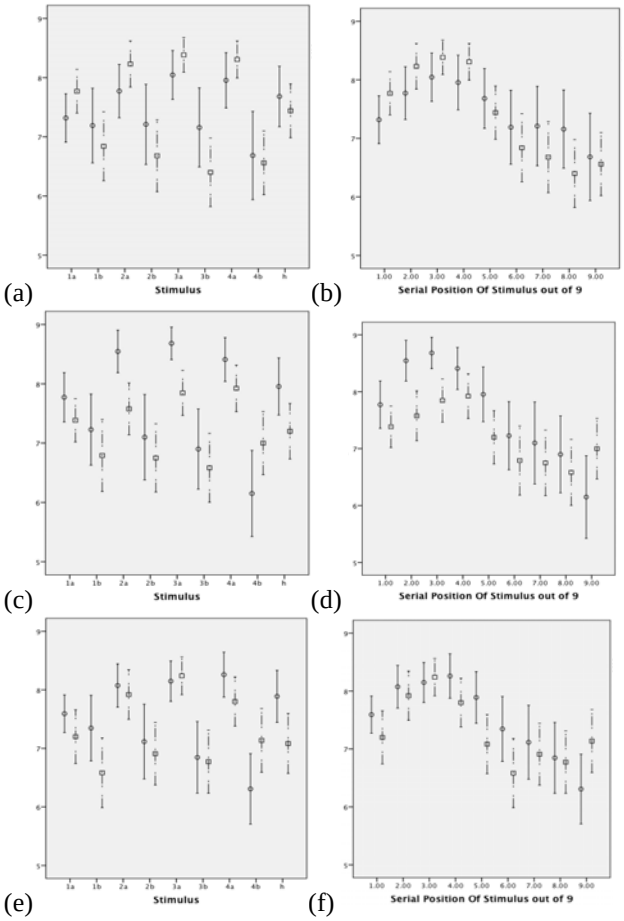

Figure 3. Error bar plots for concentration ratings by stimulus.
(a) ME by stimulus, (b) ME by serial order, (c) MS by stimulus, (d) MS by serial order, (e) Musician by stimulus, (f) Musician by serial order.
Error bar = ±1SE. Solid line is A (above median) group, Dashed line is B (below median) group.

2. The ratings made by level B (below median) participants for each variable might be better explained in terms of the relationship between their rating and the absolute rating level. Inspection of all the rating pairs reported reveals that on 7 occasions the B level participants in all groups combined had confidence intervals that encompassed the scale midpoint (5) regardless of the scale item, whereas for the A level of all groups combined, this occurred only twice. We may interpret this to simply mean that the B groups had less musical experience, and were therefore less confident with their ratings, preferring to move towards the less certain, more ambivalent midpoint of the item rating scale [16].

Overall results suggest that performances 2 and 4 were the most successful in terms of expressiveness. Most importantly, apparently objective ratings of the pieces were affected by fatigue (or that fatigue/repetition affected enjoyment aspects of the music), because expressiveness and technical execution ratings drop according to serial position. This has some potentially important implications for future research, but it should also be set in the context of a task where several ratings are made for each performance of about two minutes duration.

The results demonstrate some fairly subtle distinctions among participants with different cognitive styles, and they were not always stable, but were rather influenced by fatigue effects. These data and the concentration variable rating indicated that 5 versions of the two minute piece would have been the maximums that produced responses of good reliability, but also that music systemisers seem to be privileged in their ability to make judgements about expressiveness in music, possessing with a high level of concentration compared to the low systemiser group. High music empathisers were able to concentrate consistently throughout the study compared to other groups, but not with the same intensity as the music systemisers for the first four stimuli.

The main contribution of this study, then, is that the rating of a stimulus is highly influenced by duration and number of musical items and tasks requiring completion, and that for numerous ratings of many pieces, short excerpts should be used where possible, or the rating task should be broken into blocks of about 15 minutes each (the approximate time require to rate five pieces in the present study).

The most important implication for future research is that music cognitive styles may refer to justifications for ratings rather than the actual rating magnitudes themselves. The design of our study demonstrates that these 'preference-quality' judgements are highly sensitive to psychological noise, evidenced by the greatly varied responses given to identical performances rated twice. Thus, while individual differences may have some bearing on the way computer systems of expressiveness are rated, the experimental design is critical and future research should consider it carefully.

Acknowledgments
This research is part of the Music in my Life research program supported by the Australian Research Council.

8. REFERENCES

[1] A. Friberg, R. Bresin, and J. Sundberg, "Overview of the KTH rule system for musical performance," *Advances in Cognitive Psychology,* vol. 2, pp. 145-161, 2006.

[2] S. Canazza, G. De Poli, C. Drioli, A. Roda, and A. Vidolin, "Modeling and control of expressiveness in music performance," *Proceedings of the IEEE,* vol. 92, pp. 686-701, 2004.

[3] S. Canazza, G. De Poli, A. Rodà, and A. Vidolin, "Expressiveness in music performance: analysis, models, mapping, encoding," *Structuring Music through Markup Language: Designs and Architectures, J. Steyn, Ed. IGI Global,* pp. 156-186, 2012.

[4] R. Bresin and A. Friberg, "Evaluation of computer systems for expressive music performance," in *Guide to Computing for Expressive Music Performance*, A. Kirke and E. R. Miranda, Eds., ed London: Springer-Verlag, 2013, pp. 181-203.

[5] M. Kozhevnikov, "Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style," *Psychological bulletin,* vol. 133, p. 464, 2007.

[6] S. Baron-Cohen, "The extreme male brain theory of autism," *Trends in Cognitive Sciences,* vol. 6, pp. 248-254, Jun 2002.

[7] J. Lawson, S. Baron-Cohen, and S. Wheelwright, "Empathising and systemising in adults with and without Asperger Syndrome," *Journal of Autism and Developmental Disorders,* vol. 34, pp. 301-310, Jun 2004.

[8] A. Wakabayashi, S. Baron-Cohen, T. Uchiyama, Y. Yoshida, M. Kuroda, and S. Wheelwright, "Empathizing and systemizing in adults with and without autism spectrum conditions: Cross-cultural stability," *Journal of Autism and Developmental Disorders,* vol. 37, pp. 1823-1832, Nov 2007.

[9] G. Kreutz, E. Schubert, and L. A. Mitchell, "Cognitive styles of music listening," *Music Perception,* vol. 26, pp. 57-73, 2008.

[10] G. De Poli, S. Canazza, A. Rodà, and E. Schubert, "The role of individual difference in judging expressiveness of computer assisted music performances by experts," *ACM Transactions on Applied Perception,* under review.

[11] S. Canazza, G. De Poli, and A. Roda, "How do people assess computer generated expressive music performances?," in *Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013)*, Stockholm, Sweden, 2013, pp. 353-359.

[12] N. Schenker and J. F. Gentleman, "On judging the significance of differences by examining the overlap between confidence intervals," *The American Statistician,* vol. 55, pp. 182-186, 2001.

[13] G. Cumming and S. Finch, "Inference by Eye: Confidence Intervals and How to Read Pictures of Data," *American Psychologist,* vol. 60, pp. 170-180, 2005.

[14] M. A. Boksem, T. F. Meijman, and M. M. Lorist, "Mental fatigue, motivation and action monitoring," *Biological psychology,* vol. 72, pp. 123-132, 2006.

[15] E. Schubert and G. Kreutz, "Open ended descriptions of computer assisted interpretations of musical performance: An investigation of individual differences," in *1st international workshop on computer and robotic Systems for Automatic Music Performance (SAMP14)*, Venice, Italy, submitted.

[16] K. M. Gannon and T. M. Ostrom, "How meaning is given to rating scales: The effects of response language on category activation," *Journal of Experimental Social Psychology,* vol. 32, pp. 337-360, 1996.