Improving music composition through peer feedback: experiment and preliminary results

Daniel Martín and **Benjamin Frantz** and **François Pachet**
Sony CSL Paris
{daniel.martin,pachet}@csl.sony.fr

Abstract

To which extent peer feedback can affect the quality of a music composition? How does musical experience influence the quality of a feedback during the song composition process? To answer these questions we designed and conducted an experiment in which participants compose short songs using an online lead sheet editor, are given the possibility to feedback on other participant's songs and can either accept or reject feedback on their compositions. This experiment aims at collecting quantitative data relating the intrinsic quality of songs (estimated by peer evaluation) with the nature of feedback. Preliminary results show that peer feedback can indeed improve both the quality of a song composition and the composer' satisfaction about it. Also, composers tend to prefer compositions from other musicians with similar musical experience level.

1 Introduction

Peer feedback has become an ubiquitous feature of online education systems. Peer feedback consists in letting students or participants in a class revise, assess and more generally comment on the work of other students. This model is opposed to the traditional one in which students' works are evaluated only by a teacher. Peer feedback is acknowledged to bring many benefits [Rollinson, 2005] such as saving teachers' time as well as other pedagogical positive effects [Sadler and Good, 2006]. With the increase of online learning communities and MOOCS [September, 2013], peer feedback is becoming more and more popular.

Peer feedback is not only useful in pedagogical contexts, it can be also used in creative tasks. In music composition, collaborative composition has been addressed in several studies [Donin, forthcoming 2016]. There are online creative communities in which music is composed collaboratively by several users [Settles and Dow, 2013].

In those creative contexts, the following questions are legitimate: to which extent peer-feedback can affect the quality of a musical composition? What is the influence of the musical experience of the composers involved in this process? To address these questions we have designed a music composition experiment based on anonymous one-way feedback with no dialogue. Such a scenario differs from typical collaborative composition contexts in which composers work together hand by hand in a composition. The experiment is not aimed at being realistic or to propose a new tool for collaboration composition, but specifically to collect quantitative data regarding the relation between feedback, skills and song quality.

We focus on the role of peer feedback in music composition, specifically in *lead sheet* composition. A lead sheet is a representation of a simple song consisting of a melody and a corresponding chord grid. We propose an experiment in which peer feedback consists in suggestions of changes of certain parts of the lead sheet: specific notes or groups of notes or chords. These musical suggestions can be accompanied by a text explanation. Once a feedback is posted by a participant, it can be reviewed by the composer, who then decides to either accept it (and modify the lead sheet accordingly) or discard it.

Additionally to the sheer effect of feedbacks, we also examine the characteristics of the composer, commentator or judge of the participants. Indeed, having an extended experience in music composition might be seen as a prerequisite to write a nice song or to give useful suggestions. However, previous research showed that expertise might not be as critical as we could expect [Frese *et al.*, 1999].

2 Description of the experiment

Participants are instructed to write a short composition using an on-line lead sheet editor [Martín *et al.*, 2015]. Then they are asked to give feedback to another participant's composition, and finally they are asked to improve their own original composition using feedback posted on their composition. Participants are divided randomly in two groups: participants in the control group (G1) do not receive any feedback, and try to improve the song by themselves, whereas participants from the experimental group (G2) may use the feedback received to improve their own song. The existence of these two groups is ignored by the users so that the results are not biased.

As we are trying to assess the impact of feedback on the quality of a music composition, we need to estimate the *quality* of all compositions as well as their various variations during the experiment. To do so we use social consensus to de-

termine the quality of a song: participants listen and are given the possibility to "like" other participants' compositions. The quality of a song is then simply determined by the number of likes obtained for that song. In the next section we describe in detail each phase of the experiment:

2.1 Questionnaire

Participants start the experiment by answering 15 questions about to their experience in music, and more specifically in music composition. For example, they are asked how many years they have studied music theory, how many years they have been playing in a band, which style of music they like more, how often do they compose... etc.

2.2 Original composition

Participants then write a short composition using the online lead sheet editor. A lead sheet is a particular type of music score widely used in jazz, bossa-nova and song-writing, consisting on a monophonic melody and a chord grid. All compositions have a fixed length of 8 bars; participants are not able to add or delete bars, but they can choose the tempo and the time signature of the song. Participants fill the 8 bars with a melody and chord labels (e.g. Dmaj7, Em7...etc.). Figure 1 shows a screen-shot of the lead sheet editor.

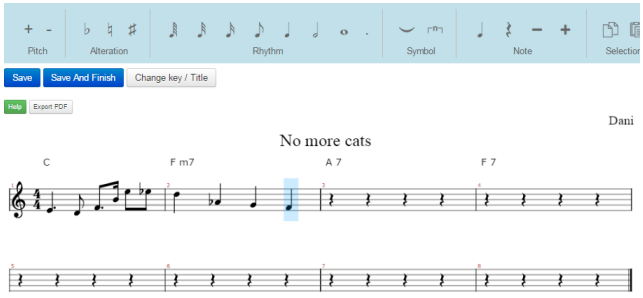

Figure 1: Screenshot of a composition being entered with the lead sheet editor.

Participants can listen to their composition with a basic MIDI player. When they are done they click on "Save and Finish". Next, they answer a questionnaire about their confidence in the quality, complexity and satisfaction on their composition.

2.3 Feedback Posting

Once they have finished their composition they are asked to give feedback to another participant by suggesting improvements in another participants' composition. Each suggestion can be at the most, two bars long. Participants can make as many suggestions as they want as long as they do not overlap. So, each participant can make a maximum of 8 suggestions (one per bar). To make a suggestion, participants must choose the bar(s) to modify, then they can change the notes and the chord symbols. Optionally, they can also leave a text comment explaining their changes. Figure 2 shows a composition in which a participant is entering suggestions with an explanation. When they are finished, they answer a short questionnaire about their confidence on the suggestions they just made as well as their opinion on the original song they modified.

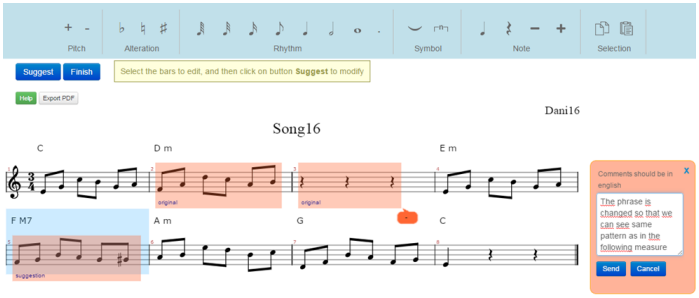

Figure 2: Screenshot showing a participant entering an explanation of the suggestion.

2.4 Improvement: Final composition

Next, participants are asked to reconsider their own composition and are asked to try to improve it. Participants from G1 (control group) are told that they unfortunately did not receive suggestions and are encouraged to try to improve their own composition by themselves. Participants from G2 see the suggestions they received from two other participants. They can listen to all the suggestions. If they like a suggestion they can *accept* it, so that it is kept and the song is automatically updated accordingly. In addition to integrating suggestions, they can modify freely their composition. Once they are finished, they answer a questionnaire about their confidence on their own improvement and on their opinion on the suggestions received.

2.5 Evaluation phase

The last step of the experiment is to evaluate pairs of compositions from other participants. Each pair of songs consist on the original song and the improved song. Participants are asked to evaluate each song by place it in a vertical display with a legend from 0 ("I don't like it") to 100 ("I like it very much"). Participants do not know which is the original and the improved song when they are evaluating. One of the versions is presented as *song A* and the other as *song B* and this assignment is performed randomly. Participants have to evaluate at least 5 pairs of songs in order to finish the experiment.

3 Results

In this section we describe in detail the results obtained from each phase of the experiment.

3.1 Population

The experiment was conducted between February and July 2015. 66 participants completed the experiment (68% men and 32% women). Mean age was 29.2 years, ranging from 19 to 61. Musical experience was measured through a questionnaire with 7 items. The scale has a satisfactory sensibility with an observed range from 7 to 41 (out of 0 to 42) and we observed a mean of 28.7 with a Standard Deviation (SD)

of 8.9. The intern consistency is satisfactory (Cronbach's alpha=.82).

Composition experience was measured through a questionnaire with 5 items. The results show an overall low level of experience concerning composition in our sample with a mean 6.9 (SD=6.1) on a scale ranging from 0 to 30). The intern consistency is satisfactory (Cronbach's alpha=.85).

3.2 Composition effects

Each participant was randomly assigned to either the control group (G1) or the experimental group (G2). No significant differences were observed between the two groups in relation to age, gender, musical experience or composition experience.

Composition evaluations

During the evaluation step, we checked if participants had listened to the songs before evaluating them. On the 1195 evaluations made, 219 were made without listening to the song. We removed those evaluations.

The songs were evaluated by an average of 8.8 different judges. The mean score of the evaluations made during the evaluation phase is 53.25 (SD = 13.26) on a scale ranging from 0 to 100. However, judges might be more or less strict, and some songs might have been evaluated by a particularly strict or generous participant. To take into account the severity of the judges, we have standardized the evaluations to get z-scores where the mean and standard deviation used are based on all the evaluations made by a given participant. As a result, the mean of the standard scores is approximately equal to zero, and a standard deviation of approximately .50. It should be noted that this final score correlates strongly with the raw score (r=.84). This result indicates that we had enough evaluations for each songs to avoid any severity bias.

Original Composition

The questionnaire that participants were asked to complete after finishing the original composition included self-estimation questions about the quality, complexity and satisfaction for their composition on scales ranging from very bad/simple/unsatisfied (0) to very good/complex/satisfied (6). We also asked them to evaluate the time they spent to make their composition and if they used an instrument to help them to compose (and which instrument if they did).

Results show a mean quality of 2.8 (SD=1.5), a mean complexity of 1.9 (SD=1.6) and a mean satisfaction of 3.2 (SD=1.6). Only the complexity is significantly different to the center of the scales which is 3 (T(65)=-5.27 ; $p<.0001$). This means that the participants tend to judge their work as rather simple (low complexity). We also observed positive and significant correlations between these three measures, ranging from r=.41 to r=.80.

During the suggestion step, we asked the participants to also rate the quality and complexity of the songs they had to comment. Each composition from the experimental group (G2) was commented by two different participants. In the end we obtained the score from the author and two other scores from two different commentators. Interestingly, there was no correlation between the scores from the original composer and the ones from the commentators (r<.10), but the two commentators did agree together on the quality (r=.80) and on the complexity (r=.70).

Moreover, from the judgments done during the evaluation phase (in which participants evaluate pairs of songs from other participants), the measurement of the quality of each original song (standardized to z-scores) allows us to estimate the composition skills level of its author. Surprisingly, we observed that the quality of the original song is only marginally related to the composition experience (r=.18, p=.15) or to the musical experience (r=.19, p=.12).

We also asked the participants whether they used an instrument to help them in their composition. Results show a marginally significant effect in favor of the use of an instrument on the mean quality score ($T(64)$=-0.87, p=.38).

The mean duration of the composition time of the song as evaluated by the participants is 30 minutes (SD=32 min) ranging from 1 minute to 240 minutes. This evaluation is largely underestimated by the participants because the real duration calculated from the time spent on the composition software is significantly longer (m=67 min; $T(65)$=4.20, $p<.001$). The correlation between these two durations is not very high, but significant (r=.46, $p<.001$) indicating that the error of duration estimation is not exactly the same for everyone. Interestingly, we observed that the quality of the original songs (from the evaluation phase) is not linked with the time spent to compose, whether it is subjective (r=.04) or objective (r=.03). This result suggests that in a situation where there is no time constraint, the amount of time devoted to compose has no effect on its quality.

Finally, there is a difference in the consensual quality of the original song, obtained from the evaluation of several participants (0.07 in G1 vs. -0.15 in G2). This could be due to differences in the group of judges evaluating each song.

Suggestions

In the questionnaire filled after making the suggestions, participants were asked how much do they think the song they are revising will be improved due to their modifications (on a 7 points Likert scale ranging from 0 "very little", to 6 "very much").

The participants from G2, the experimental group (N=30), received two suggestions for their final composition. Once they finished, we asked them if the suggestions received were interesting (on a 7 points Likert scale ranging from 0 "very little", to 6 "very much"). Additionally, we recorded the number of suggestions they received and the number of texts comments received.

We ran a series of correlations between these measures and the improvement effect (the difference between the original song and the final song on the quality judgment score). None were significant, suggesting that neither the number of suggestions received nor the number of explanations for that suggestions have an impact on the improvement of a song.

Final composition

Overall, we can see that the control group, G1, does not improve significantly between the original song (m=.07) and the final song (m=.12) (improvement effect = .05, $T(35)$=0.94, p=.35). However, we do see a significant improvement for

the experimental group, G2, between the original song (m=-.15) and the final song (m=.08) (improvement effect = .23, $T(29)=2.47$, $p=.02$). See Figure 3.

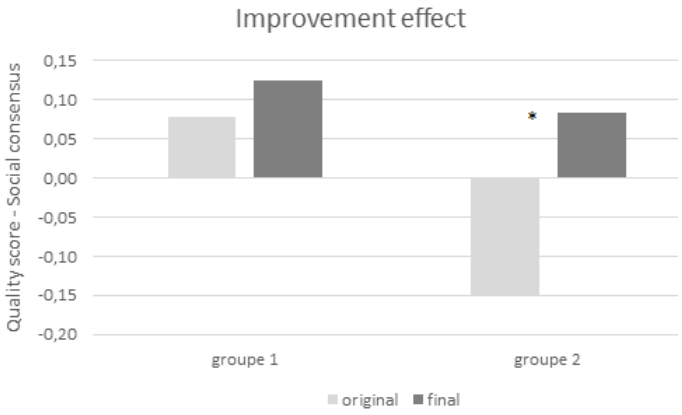

Figure 3: Difference between the original song and the final song on on the quality judgment score for the group without feedbacks (G1) and the group with feedbacks (G2).

We also examined the subjective evaluation of the participants concerning the improvement of their song. We constructed two composite scores. One from the self-evaluation scales of the original song (quality, complexity and satisfaction), one from the self-evaluation scales of the final song (quality, complexity and satisfaction). The intern consistency of those composite scores are satisfactory (the two Cronbach' alphas are above .81). We then conducted a mixed *between participants* (control and experimental groups) x *within participants* (original and final song) analysis of variance. We observed a significant interaction between groups and songs ($F(1,64) = 7.07$, $p=.01$). To explore this interaction, we used a post-hoc analysis with Tuckey HSD tests. Results show that participants who received suggestions had a significant improvement between the original and final song ($p<.001$) while the control group had no improvement ($p=.49$) See Figure 4.

When evaluating songs, users did not know which song was the original and which one was the final, as the order of the songs was determined randomly. This was a design decision to avoid the fact that participants could tend to rate better the final song, as it is supposed to be improved. Aditionally we wanted to ensure that songs were not better rated just because they had more modifications. To check this point, we used and melodic similarity algorithm [Urbano *et al.*, 2011] to estimate the similarity between each original and final songs. The correlation between the percent of similarity and the improvement effect based both on the composer's subjective opinion and on the scores from the judges are low (r=-.36, p=.003 and r=-.19, p=.13), which suggests that the improvement is not linked to the dissimilarity between the two versions.

Lead sheet editor

The software used was developed specifically for the experiment and we asked participant whether it was frustrating (0) or helpful (6) to compose with it. Results show a mean of 3.13 after the first composition and 3.41 after the final composition (the difference is not significant) which means that even if the online editor was not specially helpful, it did not hinder the composition process.

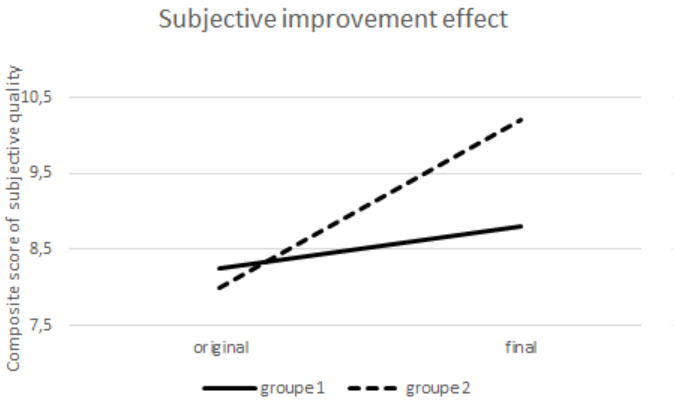

Figure 4: Self-esteemed quality of the original and final songs for the group without feedbacks (G1) and the group with feedbacks (G2).

Experience effect on evaluations

To find out whether musical experience has an impact on the way participants judge song from other participants. We divided our sample of participants in two groups according to their experience as musician (based on the median). We also divided our sample of songs according to the experience as musician of their author. We then ran a two-way ANOVA to explore the effect of the experience of the judges according to the experience of the compositor. Results show a crossed interaction between these two variables ($F(1,61)=7.63$, $p=.007$) as illustrated in figure 5. These results indicate that experienced judges give high scores to songs from experienced authors and low scores to songs from non-experienced authors. It is exactly the opposite for the non-experienced judges. This means that participants tend to prefer compositions from other participants with similar experience. This could explain the difference in the evaluation of the original songs in G1 and G2. The groups of judges evaluating each song could have different level of expertise.

4 Conclusion

The aim of this experiment was primarily to examine quantitatively the impact of peer feedback in music composition and secondly to assess how important is the experience of the participants as musicians or composers in the whole process. Before any improvement or suggestions, participants had to write their first song. Interestingly, results show that participants' previous experience in composition did not impact the quality of their song. The same pattern was also found for the participants' previous experience as a musician. These two results suggest that the quality of a song (based on social consensus) does not really tap in musicality but in something

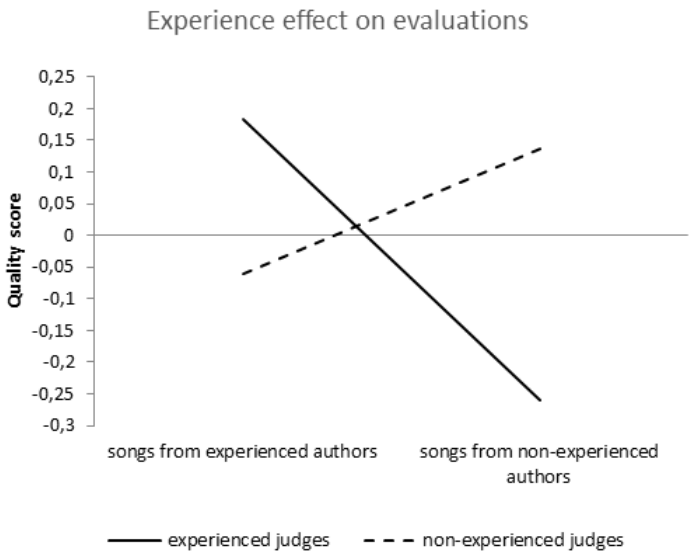

Figure 5: Interaction between the experience of the author and the experience of the judges on the quality score.

else, presumably creativity. As before, creativity might play an important role [Frese *et al.*, 1999].

Results show that composers who received feedback (G2) clearly evaluated better the improved song than the original, meaning that they were satisfied with the improvement they made. Further, the evaluation based on social consensus had a longer improvement also for G2. Hence, participants who received feedbacks not only felt that they had composed a better song after the improvement step, but they actually did. This basic finding suggests that improvements in music may be achieved even without real collaboration with dialogues and active interactions, but by simple suggestions on a single occasion.

Since there is a difference on the evaluation of the original songs between G1 and G2, we wanted to verify whether experience can make a difference when evaluating songs and we found out that participants tend to like more songs that are composed by other participants with similar musical experience.

Future work may be done by going deeper in determining the influence of the participants' experience. For example, by checking when are songs more improved, taking into account the experience of composers, commentators and judges. Further, we could assess more precisely which suggestions were actually used (or accepted) by the original composer to obtain a ranking of commentators whose suggestions are most accepted, as a measure of how good commentators they are. We could check also if suggestions from experienced commentators are more likely to be used from inexperienced composers, or whether experienced composers usually accept suggestions of other composers, and how does this affects the improvement of the song.

Acknowledgments

This work is supported by the Praise project (EU FP7 number 388770), a collaborative project funded by the European Commission under programme FP7-ICT-2011-8.

References

[Donin, forthcoming 2016] Nicolas Donin. Domesticating gesture: the collaborative creative process of florence baschet's streicherkreis for 'augmented' string quartet (2006-2008). *Eric Clarke & Mark Doffman (eds.), Creativity, Improvisation and Collaboration: Perspectives on the Performance of Contemporary Music, New York: Oxford University Press*, forthcoming 2016.

[Frese *et al.*, 1999] Michael Frese, Eric Teng, and Cees JD Wijnen. Helping to improve suggestion systems: Predictors of making suggestions in companies. *Journal of Organizational Behavior*, 20(7):1139–1155, 1999.

[Martín *et al.*, 2015] Daniel Martín, Timotée Neullas, and François Pachet. Leadsheetjs: A javascript library for online lead sheet editing. In *First International Conference on Technologies for Music Notation and Representation (TENOR)*, Paris, France, 2015.

[Rollinson, 2005] Paul Rollinson. Using peer feedback in the esl writing class. *ELT journal*, 59(1):23–30, 2005.

[Sadler and Good, 2006] Philip M Sadler and Eddie Good. The impact of self-and peer-grading on student learning. *Educational assessment*, 11(1):1–31, 2006.

[September, 2013] On September. Behind the scenes with moocs: Berklee college of musics experience developing, running, and evaluating. *CONTINUING HIGHER EDUCATION REVIEW*, 77:137, 2013.

[Settles and Dow, 2013] Burr Settles and Steven Dow. Let's get together: the formation and success of online creative collaborations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2009–2018. ACM, 2013.

[Urbano *et al.*, 2011] Julián Urbano, Juan Lloréns, Jorge Morato, and Sonia Sánchez-Cuadrado. Melodic similarity through shape similarity. In *Exploring music contents*, pages 338–355. Springer, 2011.

[Van den Berg *et al.*, 2006] Ineke Van den Berg, Wilfried Admiraal, and Albert Pilot. Designing student peer assessment in higher education: Analysis of written and oral peer feedback. *Teaching in Higher Education*, 11(2):135–147, 2006.