The creative process in lead sheet composition

Daniel Martín, François Pachet and Benjamin Frantz

Sony CSL

Abstract

Popular songs have arguably a huge impact on society. It is therefore legitimate to investigate the nature of the creative act underlying popular song composition. Ethnographic experiments in song composition are difficult to conduct. This chapter describes an experiment addressing the role of *feedback* in the lead sheet composition process. 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? Participants compose short songs using an online lead sheet editor, and are given the possibility to feedback on other participants' songs. Feedbacks can either be accepted or rejected in a later step. Quantitative data is collected from this experiment that can be used to estimate the relation between the intrinsic quality of songs (estimated by peer evaluation) and the nature of feedback. 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.

The creative process in lead sheet composition

Writing a good song

Songs invade our daily lives, to such an extent that musical taste is now considered as a trait of our personality in Western societies (**?**). Writing a good song is a highly delicate endeavor. Good songs achieve a subtle balance of melody, harmony, rhythm, lyrics and sound, as well as many other factors such as the voice of the singer, the arrangements, orchestration, production, not to mention marketing and many other social factors. Some studies claim that it is possible to predict the success of a song based on objective data (editorial, acoustic, social), but these claims are highly debatable (**?**). In particular, social pressure has been shown to be a determining factor in explaining non uniform distribution of taste in our societies (**?**). As a consequence, and in spite of a recent burst in the study of the music composition process (**?**), very little is known about what are successful strategies for composing good songs.

Feedback in song writing

During the composition process of a song [1], many types of interactions take place. First, when there are several composers writing a song (usually in duos like Lennon and McCartney), they work collaboratively to exchange ideas, or to perform trial-and-error explorations. During rehearsals, performers also report feedback to composers; either explicitly, by commenting certain parts of a song or suggesting changes, or implicitly, by performing the song differently from the original composers' instructions. Later, when the composition is performed live, the audience implicitly evaluates the composition, for example by applauding effusively if they like it or less warmly if they do not. Eventually, the number of hits, or downloads from music sites also gives some sort of feedback from the very end of the creation chain.

---

[1] Although it is difficult to draw a line between experimental and popular music, we refer here to popular songs, i.e. songs that are composed with the aim of being performed publicly, distributed, sold, and more generally aiming at pleasing a specific audience.

Intuitions

Several ethnographic studies have been conducted to study the composition process for contemporary music genres: studies on collaborative composition, **?**, as well as on the influence of performers in a composition, **?**, but very little work has addressed popular music composition. The intuitions and the work described in this chapter mostly originate from the experience of the last author in song composition, in the jazz and pop genres. Many insights were collected during the song composition process taking place between 2011 and 2013, which led to the publication of two music albums, one in French pop (**?**) and one in jazz (**?**). In order to study the composition process at stake, composition involved 2 persons in each case, and all the interactions were video recorded (see, e.g., Figure **??** and Figure **??**). This huge amount of recordings has not yet been fully analyzed, but has provided various insights about the song composition process:

1. Consensus without Understanding. One of the most interesting one is the fact that although there are usually many disagreements along the way (related to the different taste of the co-composers for instance), there is always a striking moment when the song is finally right, when it grooves, where everything (melody, harmony, sounds) fits together, and when *everyone agrees* without discussion, but also without knowing *why it works*. The same feeling can be observed by crosswordists finally guessing a tricky definition: they know they found it, they do not have to check the solution, but they do not know why, i.e. there is no clear validation process (contrarily to, e.g., a math problem, which usually contains a validation procedure).

2. Creating versus Evaluating. One of the striking force of co-composition is that the composition dialog enables the actors to switch between 2 mental modes: a creative mode and an evaluation mode. The creating mode is needed when there is a difficulty and a solution has to be found: one of the co-composers has to find a way out. The evaluation mode consists in revisiting the other co-composer's

idea and correct it, changing it or challenge it in some way. When composition is conducted alone, the composer has to switch constantly between these 2 modes, which is cognitively the most challenging aspect of music composition (and arguably of creation in general).

3. The importance of lead sheets. In jazz and in pop music, lead sheets play a major role. Of course some pop song composer do not write explicitly lead sheets. For instance, Paul McCartney always claimed he did not know how to write music. Same for Django Reinhardt, composer of the famous tune *Nuages*. However, even when they are implicit, lead sheets are the primary form of a song creation: a melody with a chord grid. The composition process can be envisaged, at least in first approximation as being lead sheet based.

However interesting, the analysis of such composition dialogues do not provide quantitative data about the strategies involved.

All interactions happening during the composition process can be defined as different forms of feedback. This is the motivation behind the experiment described here, in which we assess the impact of feedback in the quality of a music composition.

The concept of feedback, and more concretely, peer-feedback, which refers to feedback provided between equals, has become popular due to the increase of e-learning systems and online social networking sites, **?**. In such sites, internet users exchange ideas about a given subject., e.g. music composition, **?**, or music production, **?**. In these sites, users collaborate to compose and produce music respectively. Also, online courses like MOOCS, are becoming more and more popular, **?**.

Similarly, in the pedagogical domain, new teaching methods are emerging in which students receive feedback from their peers, rather than only from the teacher. Peer-feedback has been proved to bring several benefits in education. **?** shows that peers can provide useful feedback, of a different nature than the ones that a teacher can provide, (e.g. more informal), and that peer-feedback enforces collaboration between students and helps them become more critical. **?** state that peer-grading reports many benefits when students are trained by the teacher in the skills of grading. Similarly, **?**

shows the benefits on training students for peer-reviewing in the context of an ESL (English as Second Language) class.

In order to evaluate the impact of peer-feedback on the *quality* of a song, we restricted our study to a specific type of feedback. In this experiment, feedback is provided in a similar way to which *corrections* are made in a learning context like ESL redacting: by proposing specific modifications in certain parts of the whole work and possibly commenting them. Even though text correction and music composition reviewing are very different tasks, they are both a way of providing feedback. Therefore, in this experiment, feedback is suggested by peers in the form of modifications of certain parts of the composition. Feedback in this experiment is anonymous, because we analyze the quality of the musical suggestions regardless of the relationship between composers and commentators, as it is proven that relationships between peers can influence collaborative music composition, **?**.

Lead sheets

This chapter focuses on lead sheets. A lead sheet consists of a melody, most of the times monophonic as it is usually intended to be sung, and a sequence of chord symbols representing the harmony (e.g. Amaj7, Dm7, E7b9...etc.). Lead sheets are then arranged, orchestrated and more generally realized and produced. Such a viewpoint is of course reductionist, as it is probable that dimensions that are not represented in lead sheets, such as sound, instrument or voice timbre, do influence the composition process (albeit in still unknown ways). However, we restrict our experiment to lead sheet as they are a primary form of pop song composition, and have an acknowledged existence [2].

Lead sheets are widely used in pop music, as well as in jazz, bossa-nova and many other popular music genres. Figure **??** shows the lead sheet of *Pretty Late*, a composition from **?**. Jazz standards are usually played in jam sessions, where musicians can play together with other musicians without knowing each other. Musicians use the

[2] lead sheet, for instance, are the primary assets of music publishing companies, which is a tangible sign that they somehow represent the essence of a song, regardless of its possible interpretations

lead sheet as a guide. A typical jazz ensemble is composed by a musician playing the melody (e.g. a saxophone or a trumpet), another (usually a guitarist or pianist) playing the harmony defined by the chord grid, a bassist, also following the chord grid, and a drummer. Lead sheets are suitable for a music genre like jazz, in which musicians can play in a very free way. Even though the lead sheet represents only the essential information of a song, jazz musicians have enough knowledge to play the song by following it, even if they do not know the song. E.g. a guitar player infers what chords to play from the lead sheet's chord grid. Then, he decides other dimensions of performance, such as which notes to play, where in the fingerboard, the tempo, etc.

The experiment we describe below is based on an online lead sheet editor (**?**) that was used in particular for populating a large lead sheet database **?**.

Experiment on feedback in lead sheet composition

To which extent peer-feedback can affect the quality of a music composition? How does musical experience influence the quality of feedback during the composition process? To answer these questions we propose an experiment in which participants compose songs, provide feedback to other participants and try to improve their own initial composition.

Participants are divided randomly in two groups: the control group (G1) does not receive feedback, so participants from this group have no external help when improving the composition. Participants from the experimental group (G2) receive feedback from two other participants and can use it to improving their composition.

To measure the quality of a music composition we take into account two types of quality: the *subjective quality* is provided by the composer of the song. The *consensual quality* is obtained by social consensus, i.e. by aggregating the opinions of several participants. Further, we estimate a composition experience level of each participant, as well as a more general musical level, by asking them to fill a questionnaire before starting the experiment.

In the next section, we describe each step of the experiment.

Experiment protocol

The experiment is performed online, participants are recruited from mailing lists of composition in jazz and pop music. We describe now each experiment phase:

Experience questionnaire. When participants are logged in the online tool, they are asked to fill a questionnaire about their music skills in performing and composing. 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, or how often do they compose (see Figure **??**).

Song composition. In a first phase, participants have to compose an 8 bar lead sheet using the online editor. Participants cannot add or delete bars. However, they are free to choose the time signature, the tonality and the tempo. Participants have to enter the melody and the chord grid with chord symbols (e.g. Cmaj7, Dm...etc.). Participants can listen to their composition with a basic MIDI player. Once they are done they are cannot edit the song anymore. Next, they answer a questionnaire about their confidence in the quality, complexity and satisfaction on their composition.

Providing feedback. In this phase, each participant is assigned randomly to another participant's composition, and is asked to make suggestions to improve it. These suggestions are expressed as modifications of notes or chords symbols on a specific region of the composition with a duration of one or two bars. As many suggestions as wanted can be issued as long as they do not overlap with each other. 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 write a text comment explaining their changes. See Figure **??**.

Song improvement. In this phase of the experiments, participants are asked to improve their initial song. Those from the control group (G1) try to improve it by themselves with no help from peers, whereas those from the experimental group (G2) are invited to review the suggestions from other participants, play them and accept or reject them. They can also modify freely their song. At this point, each participant has produced two versions the song: the original and the final one. Figures **??** show an

example of an original composition from the experiment and **??** the same composition after improvement. Once they are finished, they answer a questionnaire about their confidence on their own improvement and on their opinion on the suggestions received.

Evaluation. In the final phase of the experiment, participants must evaluate at least 5 pairs of songs of other participants by listening to them and giving them a note between 0 and 100 according to how much they like it. Participants do not know which is the original and the improved song. One of the versions is presented as *song A* and the other as *song B* and this assignment is performed randomly. Each pair of songs is presented separately, through a pagination system designed so so that participants are forced to listen and evaluate songs in a short amount of time. We want to avoid participants to listen to a song once and then get back to it another day to listen to it again and evaluate it, because this could bias the results, as previous research has proved that previous exposure to a melody has an effect on preference: **?**.

Results

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

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. Internal 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). Internal consistency is satisfactory (Cronbach's alpha=.85).

Composition effects

Each participant was randomly assigned to either the control group (G1) or the experimental group (G2). **?** show that socio-demographic factors influence musical skills. However, no significant differences were observed between the two groups in relation to age and gender, nor on musical experience nor composition experience.

Composition evaluations. During the evaluation step, we checked if participants had actually 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 disparity in the judge ranking schemes, we 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 the standard deviation is 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 bias due to the disparity of the judges.

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, with 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 a musical 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 phase, 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 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 consensual 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 time taken to compose 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: 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 the same for everyone. Interestingly, we observed that the quality of the original songs (from the evaluation phase) is not related with the time spent to compose, whether it is subjective ($r = .04$) or consensual ($r = .03$). 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, however, be explained by 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, the number of suggestions they used 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.

The suggestions received could concern the notes or the chords, and the receiver of the suggestions could choose to use them or not. Based on the 154 suggestions that were made, 38 concerned both notes and chords, 57 only notes, 44 only chords and 15 were only text comments.

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 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 **??**.

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), and one from the self-evaluation scales of the final song (quality, complexity and satisfaction). The internal 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 **??**).

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. Additionally we wanted to ensure that songs were not better rated just because they had more modifications. To check this point, we used a *melodic similarity algorithm* from **?** to estimate the similarity between original and final songs. The correlation between the amount 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 related to the dissimilarity between the two versions.

We also see no relation between the subjective improvement and the number of suggestions received ($r = .07$). However, a significant correlation appears when we look at the number of suggestions used ($r = .47$). This result suggests that the subjects proceeded to a selection of the suggestions provided and did not simply integrate them. Due to the limited size of the sample, we took extra care when looking at the scatter plot to ensure the absence of atypical subjects.

Based on the 154 suggestions received by the subjects of group 2, a multiple linear regression was calculated to predict the subjective improvement effect based on (a) the subjective quality of the original song, (b) whether the suggestion concerns notes, (c)

whether the suggestion concerns chords and (d) the number of suggestion used. A significant regression equation was found ($F(4, 149) = 29.21$, $p<.001$) with an $R^2 = .44$). As expected, the subjective quality of the original song is the strongest predictor (β=.57, $p<.001$). Also as expected, the number of integrated suggestion is a significant predictor ($\beta = .17$, $p<.001$).

Interesting results were obtained concerning the type of suggestions. When suggestions concern chords, they are a significant predictor, with the same weight as the number of integrated suggestions (β=.19, $p<.001$), but it is not the case when the suggestions concern notes ((β=-.01, p=.92). This is interesting because it suggests that improvement of songs are made through chords modification rather than melodic changes. Confirming our previous results, adding the composition experience level or music experience level of the commentator to the equation appears to have no effect on the improvement and does not change the equation (β=.07, p=.38 and β=.11, p=.14, respectively).

Lead sheet editor. The software used was developed specifically for the experiment and we asked participants 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.

Experience effect on evaluations. We assumed that the consensual quality is a reliable evaluation. Similarly, **?** relied on *consensual assessment* to measure creativity of children's musical compositions. Nevertheless, he found out that composers evaluated differently from other groups such as music teachers, music theorists...etc. In our case, we asked ourselves whether musical experience has an impact on the way participants judge songs from other participants, so we divided our sample of participants in two groups according to their experience as a 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 **??**. 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, which 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.

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 songs. The same pattern was also found concerning 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 else, presumably creativity (**?**).

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 finding suggests that improvements in a music composition may be achieved even without real collaboration with dialogues and active interactions, but by simple suggestions on a single occasions.

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 prefer songs composed by other

participants with similar musical experience.

Future work may be determine better the influence of the participants' experience, specially regarding consensual quality. This could be done by checking when are songs more improved, taking into account the experience of composers, commentators and judges. Further, we could check 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 this may affect the improvement of the song.

Studies such as described in **?** suggest that there some musical skills may be obtained from mere exposure to music, with no need of training. So, in addition to regarding to musical experience and composition experience of participants as evaluators we could check their *listening experience* as well.

Another question needs to be addressed: what is the impact of peer-feedback in creativity (as opposed to quality). Some studies have attempted to measure creative thinking in music, **?**. Others studies, conducted in the field of design, show that previous exposure to ideas have an effect on the creation of designs. **?** show that exposure to previous examples can have negative effects such as constraining the generation of ideas. **?** point out that experienced designers use exposed examples better than novices in their creations. Following these investigations, we could evaluate how exposure to musical suggestions influence the creativity of a composer, i.e. not only the quality but also the originality of his or her composition.

Acknowledgments

This research was conducted within the Flow-Machines project funded by the ERC under the European Union's 7th Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 291156, and by the Praise project (EU FP7 number 388770), a collaborative project funded by the European Commission under program FP7-ICT-2011-8.

Figure 1. Interaction taking place during the song composition process, leading to the album **?**.

Figure 2. Interaction taking place during the song composition process, leading to the album **?**.

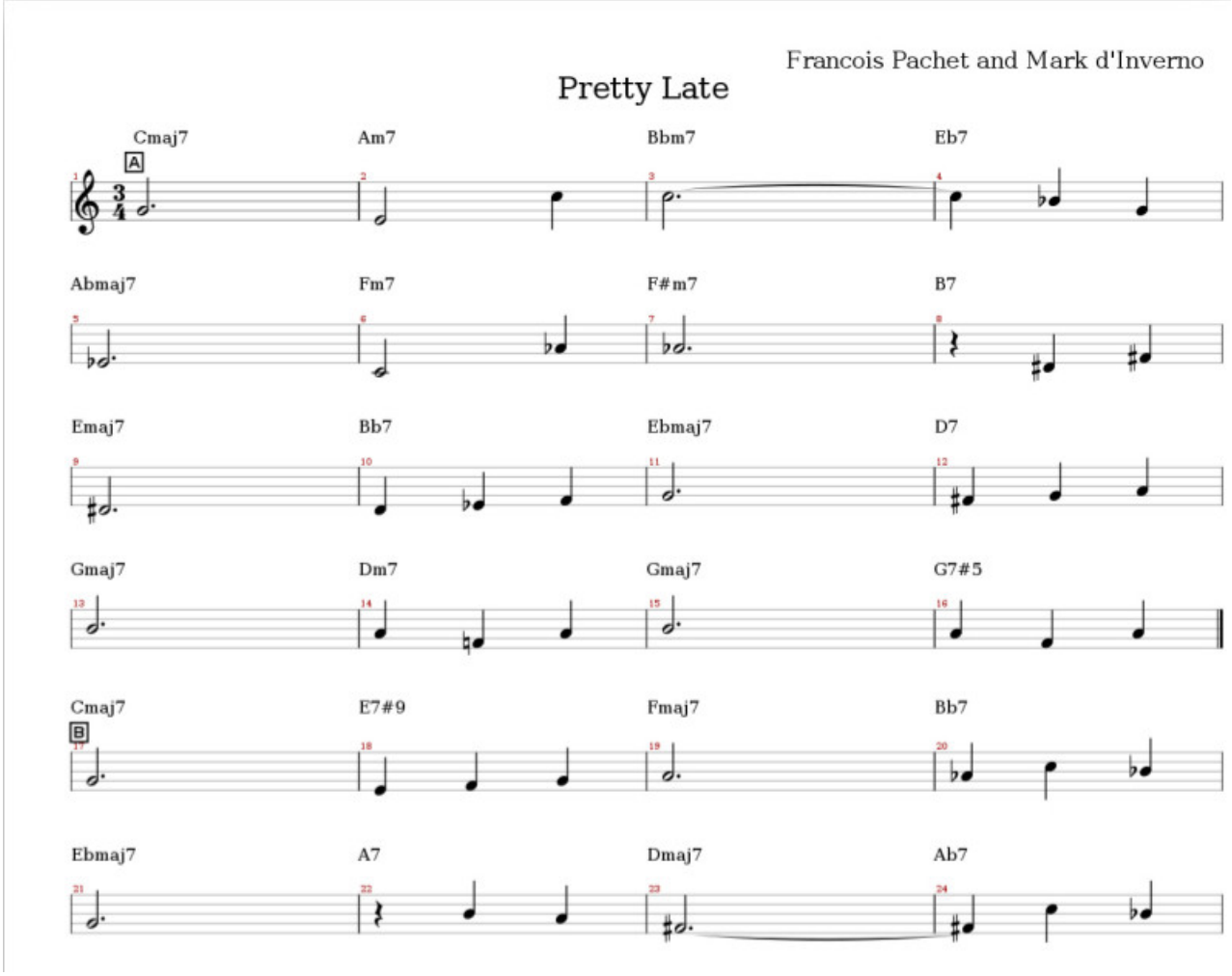

Figure 3. *Pretty Late*, a composition from **?**.

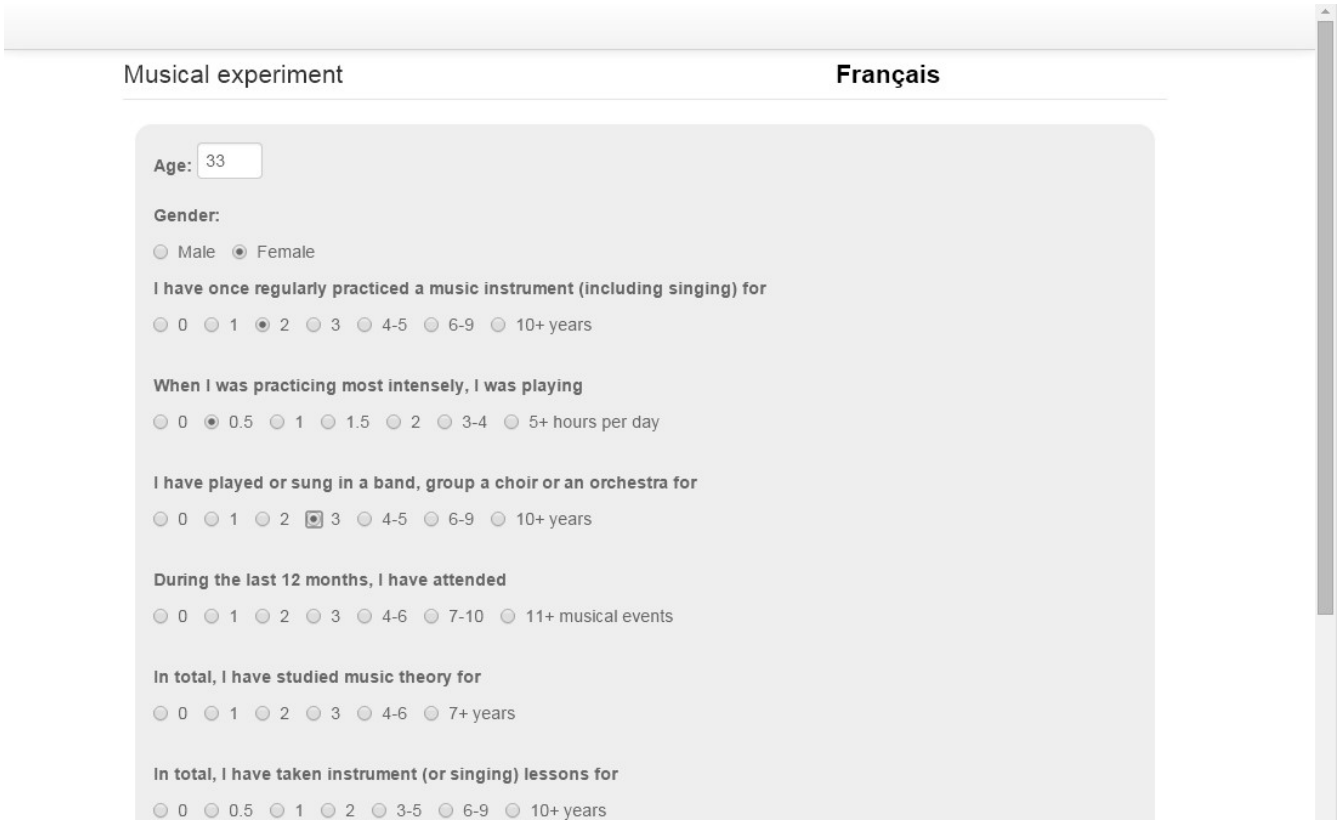

Figure 4. Questionnaire filled by participants.

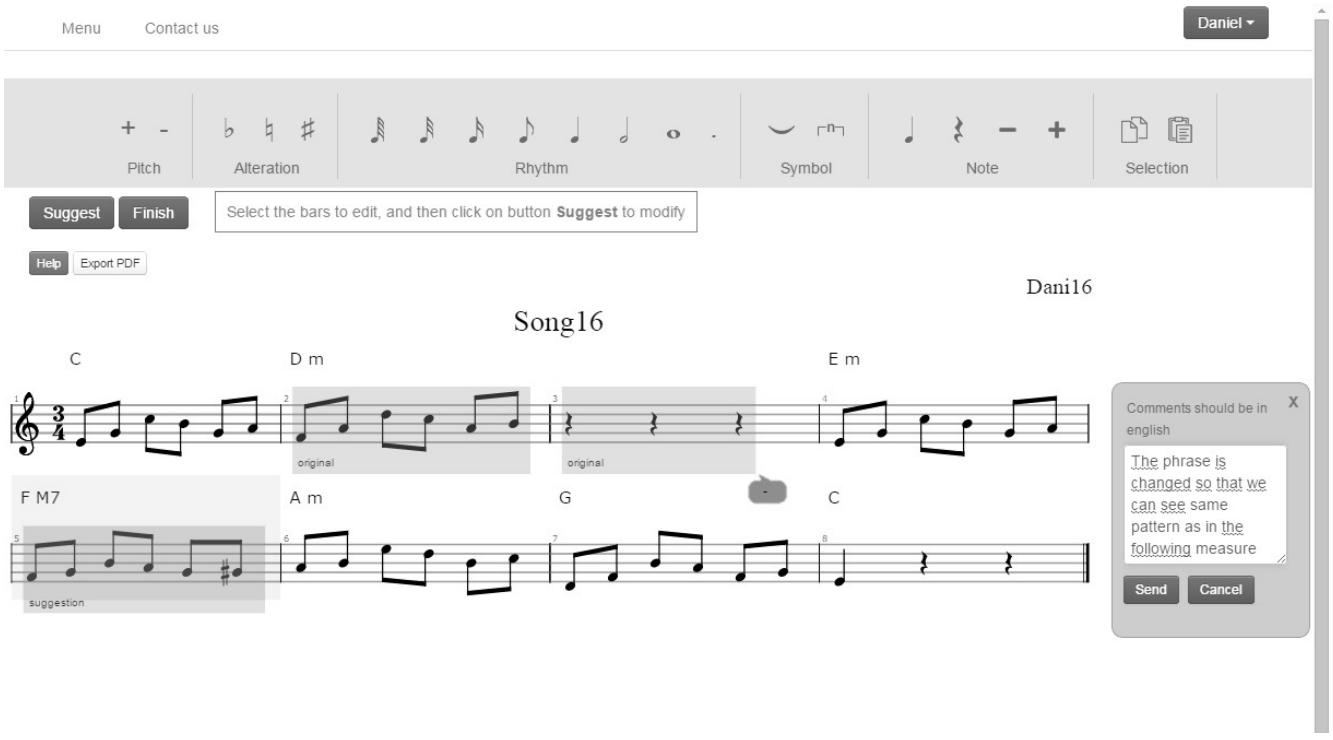

Figure 5. Example of feedback provided by a participant.

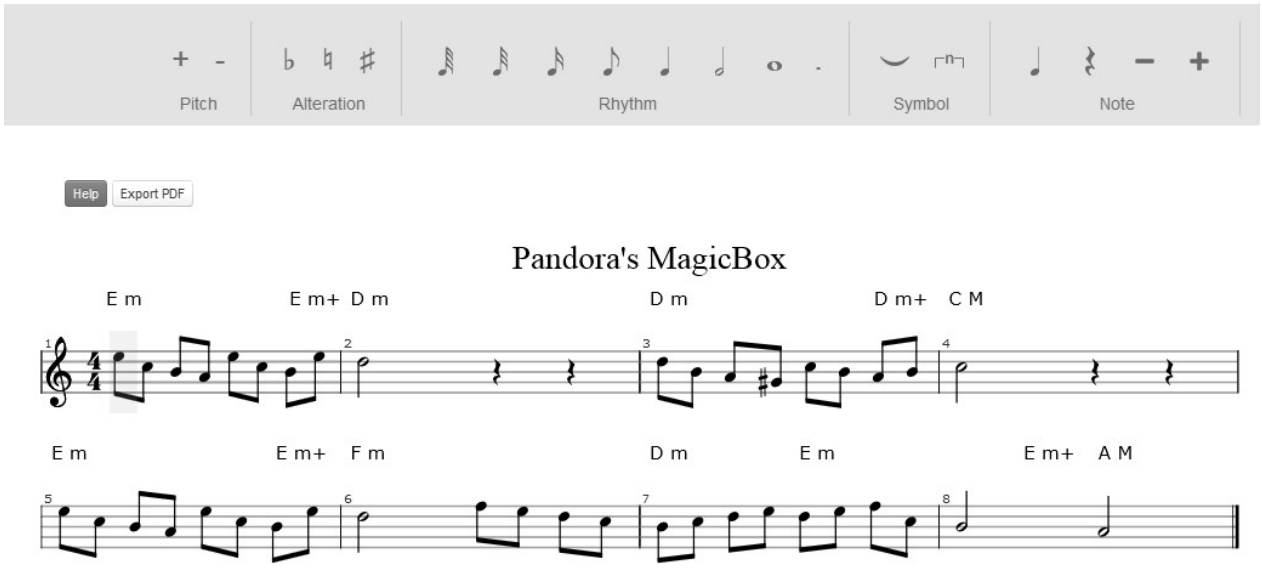

Figure 6. Example of an original composition from a participant.

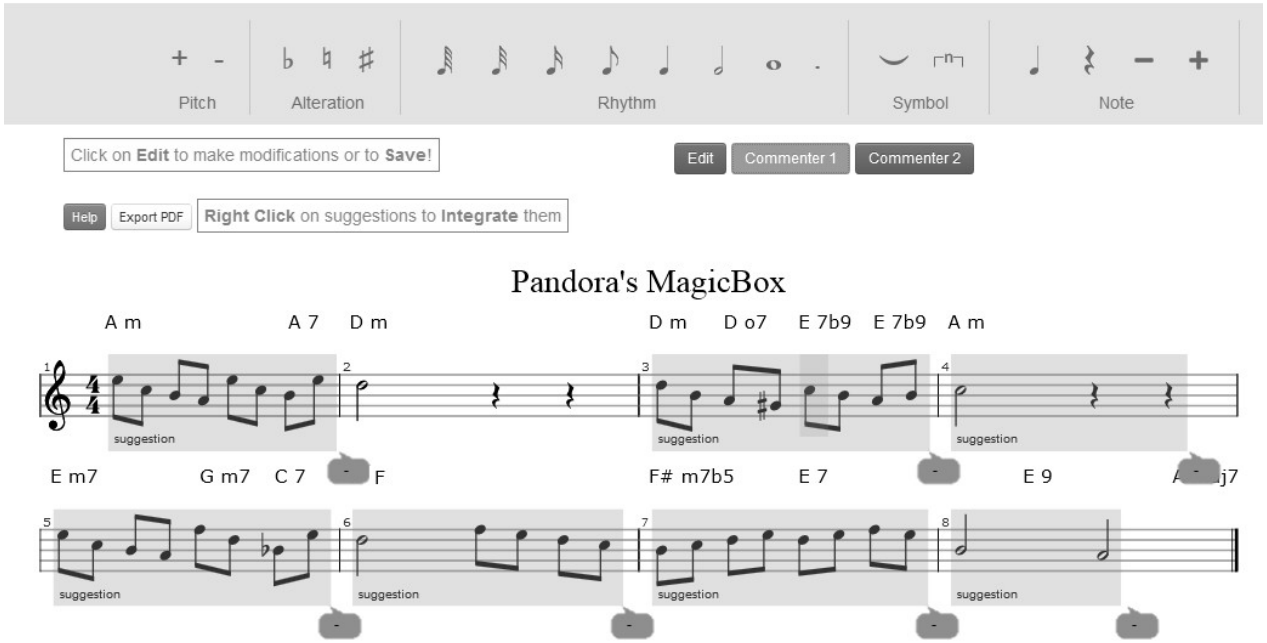

Figure 7. Example of an improved composition.

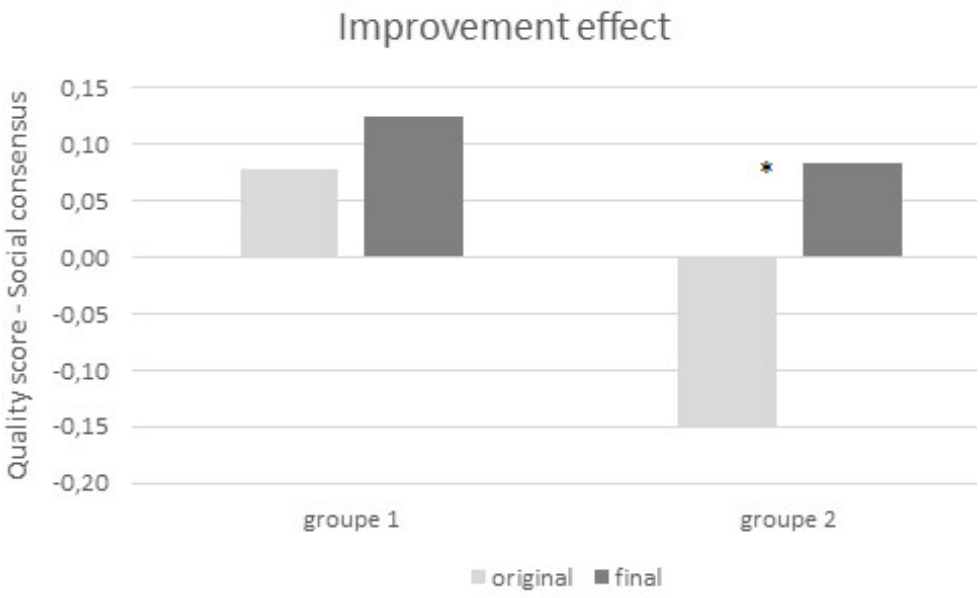

Figure 8. Difference between the original song and the final song on the quality judgment score for the group without feedbacks (G1) and the group with feedbacks (G2).

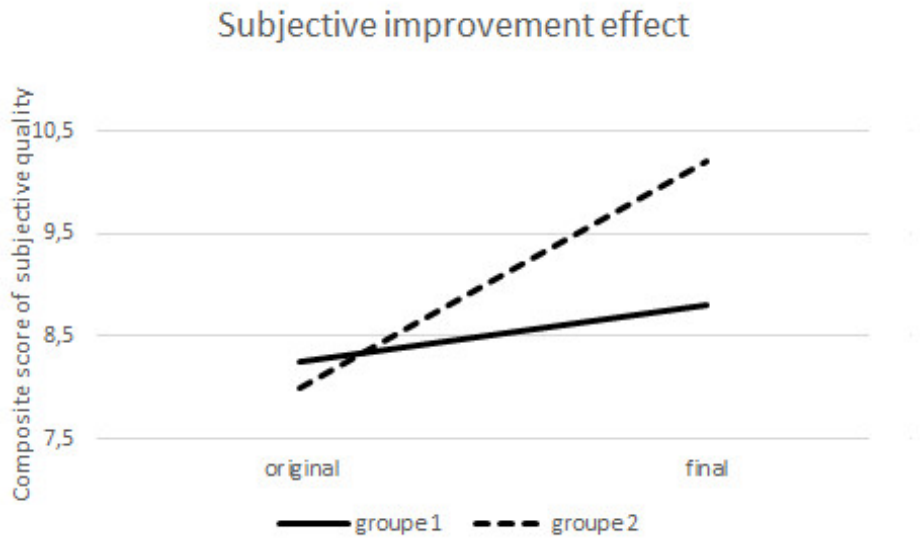

Figure 9. Self-esteemed quality of the original and final songs for the group without feedbacks (G1) and the group with feedbacks (G2).

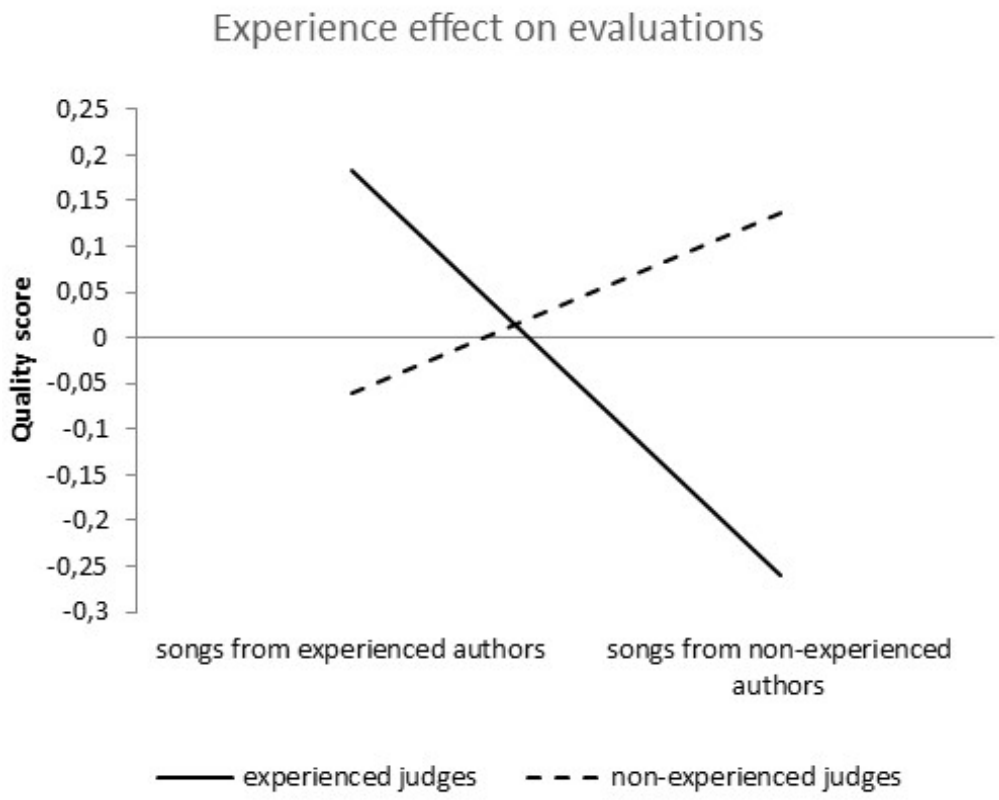

Figure 10. Interaction between the experience of the author and the experience of the judges on the quality score