

TROMPA

TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 6.9

Final evaluation

Grant Agreement nr	770376
Project runtime	May 2018 - April 2021
Document Reference	TR-D6.9-Final Evaluation
Work Package	WP6 - End User Pilots
Deliverable Type	Report
Dissemination Level	PU- Public
Document due date	30 April 2021
Date of submission	9 May 2021
Leader	TUD
Contact Person	Cynthia Liem (c.c.s.liem@tudelft.nl)
Authors	Cynthia Liem (TUD), David Baker (GOLD), Ioannis Petros Samiotis (TUD), David Weigl (MDW), Nicolás Gutiérrez (UPF), Juan Sebastián Gómez (UPF), Maria Pilar Pascual (UPF), Emilia Gómez (UPF)
Reviewers	Aggelos Gkiokas (UPF)

Executive Summary

After three years of research and prototyping, the TROMPA project is concluding. As final large deliverable, this document presents the final evaluation outcomes of the user-facing evaluations, conducted on the final prototypes of TROMPA's five use cases.

The evaluation strategies follow up on lessons learned and strategies chosen during the mid-term evaluation, reported in **Deliverable 6.8 - Mid-term evaluation**¹; broader descriptions of the prototypes in question can be found in:

- ❖ **Deliverable 6.3-2 - Working prototype for scholars**²;
- ❖ **Deliverable 6.4-2 - Working prototype for orchestras**³;
- ❖ **Deliverable 6.5-2 - Working prototype for instrument players**⁴;
- ❖ **Deliverable 6.6-2 - Working prototype for singers**⁵;
- ❖ **Deliverable 6.7-2 - Working prototype for music enthusiasts**⁶.

With regard to the **music scholars**, two evaluation studies were conducted. The first considered the evaluation of the Digital Score Edition software component in the context of the music of Gustav Mahler, and was designed in collaboration with Mahler scholar Dr Paul Banks. An online guided questionnaire paired with open-ended feedback, focusing on both the user experience and user interaction with the software, was conducted with a group of international music scholars. The users were guided to interact with specific features of the software, after which they rated the extent to which they either agreed or disagreed with a statement about the feature of interest on a 7-point Likert scale, while also having the opportunity for additional open-ended feedback.

Besides the Mahler study, another evaluation study was held on early vocal music. This study was conducted to collect user feedback on two further provisioned features relevant to music scholars: the possibility to conduct a search query that allowed users to view any score linked through the TROMPA contributor environment, and the ability to use F-TEMPO⁷ to perform partial match searching. Again, for each feature of interest, participating scholars were invited to interact with the feature, give Likert-scale ratings to statements about the feature, plus the possibility to add additional open-ended feedback.

Almost all of the participants in the user study expressed a high degree of enthusiasm about the potential for what the digital score edition component can do, and saw clear benefits of the functionality for their practice. At the same time, several participants raised specific issues with aspects of the user experience and user interface. For future work, the first and foremost changes to make to the current DSE prototype are with on-screen presentation and ease of use. For the software to be adopted widely, it needs to function more as individuals expect it to, which in turn demands more consideration of graphic design, which was considered beyond scope for the present prototype.

¹ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

² https://trompamusic.eu/deliverables/TR-D6.3-Working_Prototype_for_Scholars_v2.pdf

³ https://trompamusic.eu/deliverables/TR-D6.4-Working_Prototype_for_Orchestras_v2.pdf

⁴ https://trompamusic.eu/deliverables/TR-D6.5-Working_Prototype_for_Instrument_Players_v2.pdf

⁵ https://trompamusic.eu/deliverables/TR-D6.6-Working_Prototype_for_Singers_v2.pdf

⁶ https://trompamusic.eu/deliverables/TR-D6.7-Working_Prototype_for_Music_Enthusiasts_v2.pdf

⁷ <http://f-tempo.org>

For the **orchestras**, three evaluation studies are reported, that all were held online with members of student and youth orchestras based in The Netherlands.

The first reported user study in the orchestra use case considers a live campaign and transcription task design study. It was the first time in which orchestra members could interact with a live campaign for clef, time signature and key signature recognition crowd tasks; next to this, feedback was solicited on how to improve task design for transcription-oriented crowd tasks. Participants found the system very simple to use, possibly even too simple for their own musical expertise. With regard to transcription task design, two options were offered to the participants: (A) first encode generic notes and rests, followed by pitch correction and note duration tasks (or the other way around) or (B) encode notes with pitch within the same task, followed by note duration correction. Participants preferred option (A), but came with an even better suggestion, that ultimately was taken forward in our further studies: first transcribe the rhythm of a given segment, to be followed by pitch correction.

The second study was run in several workshop sessions, in which participants were shortly introduced to TROMPA, after which they were invited to silently work for an hour on a collaborative transcription of segments from the Beethoven Wind Sextet, op. 71, where the amount of segments was adjusted to the participant population size. For this collaborative transcription, the participants would go through a live campaign with subsequent stages of clef recognition, time signature recognition, key signature recognition, rhythm transcription, and pitch correction. With more complex transcription tasks added, the system became more difficult to use. This both reflects in longer completion times for rhythm and note transcription, as well as in the post-study system usability questionnaire responses, in which the system was not deemed as easy to use as in previous studies, that did not yet include live transcription tasks. Furthermore, tasks were considered less self-explanatory, and the interaction design of the rhythm transcription task was especially deemed very cumbersome by the participants.

One of the orchestras in this second study, the digitally-minded Almere Youth Symphony Orchestra (AJSO), was extremely enthusiastic about the prototype, and voluntarily offered to join more sessions and recruit more of their members for feedback. Taking advantage of this enthusiasm, a third evaluation study was held, in which AJSO members returned to work on improved campaign and task designs, following the feedback from the second study, and including consistency improvements, better help support, a thoroughly revised task interaction design for the rhythm transcription task, and further usability improvements on all other tasks. Here, considerably higher efficiency was measured in terms of task completion, and usability and overall user satisfaction had greatly improved.

Where the intention had been to design the tasks as campaigns which would run asynchronously, with people contributing whenever their own time would allow, due to the COVID-19 crisis, all current studies were run in synchronous sessions, to stimulate participation over a set time interval. This working form actually was appreciated by the participants, especially in case multiple members of the same ensemble would remotely sit and work together. For future campaign design, this working form could therefore have potential to further investigate, too.

Regarding the **instrument players**, two studies are reported on, that were held online and in-person (compliant with COVID-19 regulation) with professional piano majors at the mdw institute.

The first user study was fully conducted online, and served the dual purpose of (i) gaining a richer understanding of the rehearsal habits, contexts, and information requirements of our target

audience of musicians with advanced expertise in classical piano performance, and (ii) obtaining user feedback on the initially implemented CLARA prototype, to inform future development. With regard to the first purpose, the participants gave useful insight into their practice, explaining approaches to their rehearsal strategies (e.g. practicing at different tempi, capturing and studying own recordings), their rehearsal context (e.g. location and duration), the purpose of their rehearsals (e.g. focusing on specific sections), and their rehearsal recording activity. Upon the demonstration of the CLARA prototype, the ability to revisit rehearsal recordings and to navigate these through interaction with the score was universally seen as useful, as was the ability to visualise performance errors. Further individual preferences were articulated (e.g. possibility for automated page turning). At the same time, participants gave mixed responses to the potential of using a tool like CLARA in their pedagogical context, especially being concerned about additional time and effort involved.

In the second user study, participants were actively engaging with the CLARA prototype, with those who could attend in-person sessions actively performing a play-through on a digital piano. After this, participants were walked through the various features of the prototype in feature analysis mode, and then invited to freely interact with the prototype, record further rehearsal attempts, and finally, fill in an evaluation form, rating usefulness, usability, accuracy and performance of the prototype, together with open-ended feedback possibilities on each of the feature analysis component (digital score display, annotation support, tempo curve display, dynamics analysis display, error display). Finally, participants were asked to rate how likely they were to use the tool in their own future practice.

Participants were largely positive about the tool and its applicability and usefulness in the piano rehearsal context. However, the enthusiasm of several participants was moderated by limitations of the interface. The accuracy of the presented information was largely accepted, but small inconsistencies were noted. The quality of the rendered digital score was praised by all respondents, as was the utility of the tempo curve display. The tools for dynamics analysis and error visualization were largely praised. Finally, estimates of whether the participants themselves would be likely to use the tool in future ranged from neutral to enthusiastic.

With regard to the **choir singers**, a collaboration with the Cantoría vocal quartet was set up, which focuses on vocal polyphony from the Iberian Golden Age, while at the same time having a strong international network and active social media presence. Cantoría helped choosing repertoire that would ultimately be used in a participatory concert, and made professional study recordings including an instrumental track and practice tracks.

Leading up to the concert, virtual choir rehearsals were organized with the Cantamus app, developed for the choir singers prototype, with the La Violeta amateur choir. Participants were reminded of the main features of Cantamus, with special emphasis on the recording and analysis features. Then, participants were asked to learn the repertoire, record their own voice, and explore the analysis features offered by the prototype. The recordings of the singers were then converted into a virtual choir mix, combining them with the recordings from the Cantoría singers.

Then, for the participatory concert, extensive general-public advertising was done, including considerable attention in the media. Through Cantoría's contacts, the community was globally expanded to the Spanish and Latin American choral world. Registered participants were onboarded into Cantamus, and got links for accessing complementary activities (workshop about the pilot, musicological conference presentation, online rehearsal). While COVID-19 did not allow for live attendance in a participatory concert, participants were linked to a stream of Cantoría's performance. Finally, a virtual choir, based on the virtual rehearsals was synthesized.

Evaluation outcomes show that there is a need and interest among choir singers for using technologies like offered by Cantamus, that allow sustaining singing activity in the current COVID-19 situation. Furthermore, provided functionalities were well received. Several usability improvements still need to be made before the prototype would be viable as a commercial end-user solution, and participants were especially not intuitively used yet to algorithmically processed outcomes. However, there globally has been interest voiced into the prototype, further strengthening its exploitation potential.

For the **music enthusiasts**, two campaigns were run as part of the final evaluation. The first campaign unlocked music from West Africa, and was intended to evaluate the usability and workflow of the pilot in a real setting (participants using the ME platform by their own with their own devices), as well as to determine the impact of the implemented incentives (e.g. scoring system, contributors' ranking, music recommender system based on emotional content) and the quality of the annotations. Likewise, the evaluation study allowed to assess the scope of the dissemination mechanisms available, e.g. mailing lists and social networks. After logging in and going through the tutorial, participants must complete at least one of the available campaigns, with prizes being given out to the most prolific participants (as well as a randomly drawn participant).

From the first campaign, it was found that new users would initially make annotations from the initial campaigns, but would abandon the task before listening to new music from West Africa. Krippendorff's alpha was a little over 0.5 for arousal, but lower for valence and emotion. Noticing that few participants entered personal information, the workflow was improved to redirect them more explicitly to user settings. As more often with crowd campaigns, it was noticed that a small amount of users turned out very productive.

Following the insights from this campaign, a long-term campaign was formulated, which featured a daily playlist with 20 songs over a period of 27 days, and unlocked music from Latin America. By allowing for a longer time to collect annotations, the intention was to have more sustained participation. The campaign was advertised in English, Spanish, Dutch, French and Catalan through social media and official websites.

Many annotations were obtained for the tutorial songs, but less were obtained for the rest, possibly because of the absence of an external (monetary) reward. Again, there was a skewed degree of participation, with a smaller amount of participants generating most of the annotations. Participants indicated they discovered new music through the campaign, although survey feedback indicated that refinements are needed to still make the recommendations more appealing to participants.

With the unexpected COVID-19 crisis, engaging audiences and running user studies had been more challenging than foreseen at the start of TROMPA. As a consequence, many of the presented studies have been conducted in smaller-scale settings than the project had originally intended.

Nonetheless, for each of the use cases, we managed getting in touch with relevant and representative user audiences, who gave very valuable feedback to our work. With digital innovation not necessarily having been embraced yet in the classical music communities (even though the COVID-19 crisis did push in favor of this), it is important to identify and stimulate champions in leadership positions: both in the orchestras use case (with the AJSO orchestra) and the choir singers use case (with La Violeta), the most engaged and enthusiastic ensembles had very enthusiastic directors, who helped engaging the ensembles.

Many participants in our use case evaluation studies reacted enthusiastically to provided functionality, and did see future promise in our prototypes. Therefore, beyond the lifetime of the TROMPA project, it will be worthwhile to further develop and improve the prototypes. As soon as circumstances will have normalized after the crisis, it will be interesting to revisit the user studies under more ecological conditions.

Version Log

#	Date	Description
v0.1	April 28, 2021	First review version sent out
v0.2	April 29, 2021	Review commentary added
v0.3	May 3, 2021	Review comments addressed
v0.4	May 7, 2021	Minor changes
v1.0	May 9, 2021	Final version

Table of Contents

1. Introduction	11
2. Music Scholars	13
2.1. Digital Score Edition: Mahler Use Case	13
2.1.1. Aim of the evaluation study	13
2.1.2. Participants	13
2.1.2.1 Recruitment strategies	13
2.1.2.2 Participant characteristics	14
2.1.3. Study protocol	14
2.1.4. Study evaluation outcomes	15
2.2. Digital Score Edition: Early Vocal Music Use Case	18
2.2.1. Aim of the evaluation study	18
2.2.2. Participants	19
2.2.2.1 Recruitment strategies	19
2.2.2.2 Participant characteristics	19
2.2.3. Study protocol	19
2.2.4. Study evaluation outcomes	19
2.2.5 Joint study synthesis	22
3. Orchestras	25
3.1. Third user study: live campaign and transcription task design	25
3.1.1. Aim of the evaluation study	25
3.1.2. Participants	27
3.1.2.1 Recruitment strategies	27
3.1.2.2 Participant characteristics	27
3.1.3. Study protocol	28
3.1.4. Study evaluation outcomes	29
3.2. Prototype evaluation sessions	30
3.2.1. Aim of the evaluation study	32
3.2.2. Participants	32
3.2.2.1 Recruitment strategies	32
3.2.2.2 Participant characteristics	33
3.2.3. Study protocol	34
3.2.4. Study evaluation outcomes	35
3.3. Final iterations with AJSO	38
3.3.1. Aim of the evaluation study	40
3.3.2. Participants	41
3.3.2.1 Recruitment strategies	41
3.3.2.2 Participant characteristics	41
3.3.3. Study protocol	41

3.2.4. Study evaluation outcomes	41
4. Instrument players	45
4.1. Structured interview on digital piano rehearsal	46
4.1.1. Aim of the evaluation study	46
4.1.2. Participants	46
4.1.2.1 Recruitment strategies	46
4.1.2.2 Participant characteristics	47
4.1.3. Study protocol	47
4.1.4. Study evaluation outcomes	48
4.2. Interactive evaluation of the prototype	51
4.2.1. Aim of the evaluation study	51
4.2.2. Participants	52
4.2.2.1 Recruitment strategies	52
4.2.2.2 Participant characteristics	52
4.2.3. Study protocol	52
4.2.4. Study evaluation outcomes	53
5. Choir singers	54
5.1. Pilot Evaluation within the context of Renaissance repertoire	57
5.1.1. Aim of the evaluation study	57
5.1.2. Participants	57
5.1.2.1 Recruitment strategies	57
5.1.2.2 Participant characteristics	58
5.1.3. Study protocol	58
5.1.4. Study evaluation outcomes	59
5.2. Live/on-line participatory concert with Cantoría	59
5.2.1. Aim of the evaluation study	59
5.2.2 Additional functionalities of the Choir singing Pilot	59
5.2.3. Participants	60
5.2.3.1 Recruitment strategies	60
5.2.3.2 Participant characteristics	62
5.2.4. Study protocol	64
5.2.5. Study evaluation outcomes	67
5.3. Conclusions	69
6. Music enthusiasts	70
6.1 Second contest: Music from West Africa	70
6.1.1. Aim of the evaluation study	70
6.1.2. Participants	71
6.1.2.1 Recruitment strategies	71
6.1.2.2 Participant characteristics	71
6.1.3. Study protocol	72
6.1.4. Study evaluation outcomes	72

6.2 Long-term campaign: Music from Latin America	73
6.2.1. Aim of the evaluation study	73
6.2.2. Participants	74
6.2.2.1 Recruitment strategies	74
6.2.2.2 Participant characteristics	75
6.2.3. Study protocol	75
6.2.4. Study evaluation outcomes	76
6.3 User behavior analysis	78
7. Conclusion	80

1. Introduction

After three years of research and prototyping, the TROMPA project is concluding. As final large deliverable, this document presents the final evaluation outcomes of the user-facing evaluations, conducted on the final prototypes of TROMPA's five use cases.

The evaluation strategies follow up on lessons learned and strategies chosen during the mid-term evaluation, reported in **Deliverable 6.8 - Mid-term evaluation**⁸; broader descriptions of the prototypes in question can be found in:

- ❖ **Deliverable 6.3-2 - Working prototype for scholars**⁹;
- ❖ **Deliverable 6.4-2 - Working prototype for orchestras**¹⁰;
- ❖ **Deliverable 6.5-2 - Working prototype for instrument players**¹¹;
- ❖ **Deliverable 6.6-2 - Working prototype for singers**¹²;
- ❖ **Deliverable 6.7-2 - Working prototype for music enthusiasts**¹³.

The deliverable will visit all use cases: the Music Scholars in Chapter 2, the Orchestras in Chapter 3, the Instrumental Players in Chapter 4, the Choral Singers in Chapter 5, and the Music Enthusiasts in Chapter 6. In all cases, evaluation outcomes are reported following the same structure: for each of the evaluation studies of interest, the main aim of the study is listed, a description of participant recruitment strategies and characteristics is given, followed by the study protocol and evaluation outcomes. We conclude the deliverable in Chapter 7.

⁸ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

⁹ https://trompamusic.eu/deliverables/TR-D6.3-Working_Prototype_for_Scholars_v2.pdf

¹⁰ https://trompamusic.eu/deliverables/TR-D6.4-Working_Prototype_for_Orchestras_v2.pdf

¹¹ https://trompamusic.eu/deliverables/TR-D6.5-Working_Prototype_for_Instrument_Players_v2.pdf

¹² https://trompamusic.eu/deliverables/TR-D6.6-Working_Prototype_for_Singers_v2.pdf

¹³ https://trompamusic.eu/deliverables/TR-D6.7-Working_Prototype_for_Music_Enthusiasts_v2.pdf

2. Music Scholars

Here we report the results of the final user testing surrounding the web-based Digital Score Edition (DSE) software component developed as part of the TROMPA Music Scholars use case. Following our plans as detailed in **Deliverable 6.8 - Mid-term evaluation**¹⁴, we developed a guided questionnaire in order to assess the extent that the features developed as part of the DSE software accommodated those issues noted in our first round of user testing.

As the DSE was designed to work with any MEI encoded score, we developed two user tests to investigate the flexibility of the tools in handling various styles of scores. The first user test was designed in collaboration with Mahler Scholar Dr Paul Banks. The second user test was designed to demonstrate how the DSE can work with a variety of scores and is capable of integrating sophisticated metadata query searching. In testing the application with both groups, one of the main goals of the testing was to assess if we had developed a tool that was flexible enough to meet the needs of various types of musical genres as well as accessible enough to be usable by individuals without extensive formal musical training.

Below we detail both studies, noting that there is a substantial overlap in methodology and participants between the two user tests. Participants in the second study represent a subset of the first. We detail the protocol in Section 2.1 and only note deviations made for the Early Music User test in Section 2.2. A general summary of both studies can be found under Section 2.2.5.

2.1. Digital Score Edition: Mahler Use Case

2.1.1. Aim of the evaluation study

The aim of this study was to collect direct user feedback on the practical use of the DSE for music scholars as initially detailed in the mid-term evaluation report. Our goal was to assess several aspects of the user experience of the DSE. To do this, we used a guided questionnaire paired with open-ended feedback to collect both quantitative and qualitative data on both the user experience and user interaction with the software.

2.1.2. Participants

2.1.2.1 Recruitment strategies

Participants were recruited in months 35 and 36 of the TROMPA project. The Music Scholars team based at Goldsmiths used the official TROMPA Twitter account to help solicit interest in the study. We advertised the study as an opportunity to provide feedback for software developed as part of the TROMPA project. For the Mahler use case, we recruited any individuals who self identified as some form of music scholar. Participants were informed ahead of the study that it would take approximately one hour to complete over a video call.

¹⁴ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

2.1.2.2 Participant characteristics

The sample consisted of eight music scholars. Of the sample, three were based in the United States, two in Canada, two in Ireland, and one in Switzerland. Seven of the participants held degrees in Music (e.g. Music Theory or Musicology) and one a BA in Fine Arts.

2.1.3. Study protocol

Participants signed up to partake in the study using online booking software. After signing up, they were sent a confirmation email that included login details prepared by the technical team, a unique participant number used to separate personally identifiable information from the data they provided as a part of the study, and further details about the call.

The user study was run by a researcher on the Goldsmiths team. Upon signing into the call, participants were sent a Google Forms link that provided access to the guided questionnaire¹⁵. The participants were then guided through each of the pages of the Google Form by the researcher. During the user study, the researcher helped solve any technical issues faced by the participant and also transcribed any comments by the participant that were not able to be captured by the form.

Next we describe the procedure that each participant experienced. After providing consent and being informed of how their data will be used, participants completed a user study in five parts.

These included:

- ❖ Accessing and Logging Into the Software;
- ❖ First Impressions of the Software;
- ❖ Interacting with Pre-Existing Data in the Software;
- ❖ Creating your own Data in the Software;
- ❖ Open Discussion Feedback.

Prior to interacting with the software, participants were told that “By the end of the study, [they] will have attempted to use most of the central features of this software and should be able to give more specific feedback to what [they] like and do not like about the software.”

In order to lessen the cognitive load during the user study, the majority of the guided questions asked users to first interact with a specific feature of the software, then rate the extent that they either agreed or disagreed with a statement about the feature using a seven point Likert scale where 1 always indicated “I strongly disagree” and 7 indicated “I strongly agree”. Statements were written in both positive and negative terms to discourage participants from simply clicking down one side of the responses to expedite the process. The distribution of responses to each question is reported in Figure 2.1.

In order to replicate the user experience envisioned by the developers closely, the DSE team pre-populated the interface with a score containing annotations made by Mahler Scholar Dr. Paul Banks. These annotations served as the main data for users to interact with. The data and experiment were hosted on a server based at Pompeu Fabra University¹⁶. The entire study took approximately one hour to complete, though this time varied between participants.

¹⁵ https://drive.google.com/file/d/1rf5Wieg3Thkv_XQ0A3_9y8LUsaMM8AB/view?usp=sharing

¹⁶ <https://trompamusic.github.io/music-scholars-annotator/>

2.1.4. Study evaluation outcomes

Data from the quantitative portion of the study is reported below in Figure 2.1. The following figure plots a density distribution for each question reported. Discussion of salient and meaningful responses is provided below.

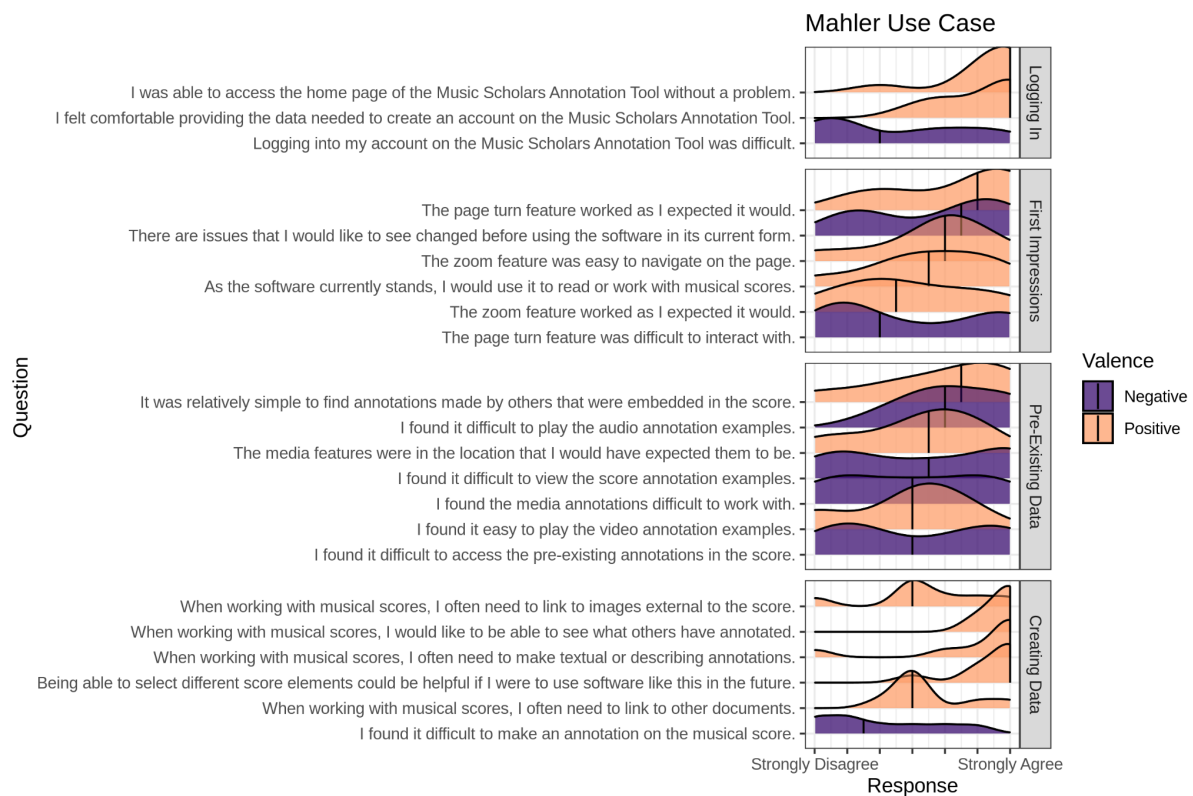


Figure 2.1. User Response Data: Mahler Use Case

In addition to the Likert survey data, participants also had the opportunity to provide free text data to provide further context surrounding their experience with the DSE. We have included quotes that were particularly relevant to our research questions as detailed in the mid-term evaluation in Table 2.1. We then contextualize these in general discussion below.

Topic	Participant feedback
Access The Software	<ul style="list-style-type: none"> ❖ Page turn button feels a bit "sticky" with delay; zoom function pushes score into left margin of the page; slightly hard to see the zoom/page turn buttons in the top of the page. ❖ I wanted to have an option to devote more, or less, of the total browser window real estate to either the score, or the annotation materials on the left hand side. I also would have liked to see a visual cue telling me where on the range of "zoom" for the score I was on...so I knew the possible range of zooming in/out and where I currently sat on that range. ❖ Right now, the turn and zoom buttons adapt to screen size, which is nice,

	<p>but it means that sometimes it gets lost. It also takes a bit of time, neither button is terribly fast or responsive.</p> <ul style="list-style-type: none"> ❖ The page turn feature worked fairly well, but the zoom feature did not. It seemed as though pages had different font sizes depending on the number of instruments in the score. A consistent font size would be preferable, I think. ❖ Expected zoom to keep score the same and not reflow/rewrite (more like an image zoomer). It would be nice to have page turn click areas at the edges of the score image rather than up top. ❖ My instinct is to use the arrows on the keyboard to navigate, which didn't work—that would be nice to have. I have tendinitis mostly from mouse clicking, so I try to prefer typing whenever possible; may be a problem some others have as well. The only thing that was mildly odd about the zoom is that because the plus/minus icons move as well, you can't just keep clicking in the same spot, have to keep chasing the zoom to go quickly. those are very minor things, though.
<p>Interacting with Pre-Existing Data</p>	<ul style="list-style-type: none"> ❖ Annotations and audio is far down on the page; coloring for annotations seems to be odd with inverse-highlighting ❖ I would vote for "table of contents" type list of all of the associated media, if there is any, that is pre-loaded for any particular score. So somewhere on the page, similar to the orange "show recordings playlist" button, either a more explicit link out to this table of contents or just a small showing of...all of the recordings, all of the images, etc., so that users can find that list of things sooner rather than having to (eventually) find the list of recordings by clicking the orange button. ❖ Audio wasn't working for me, and originally I couldn't find the text annotations. I had to scroll down a bit to find them. ❖ I am using a MacBook Air. I was not intuitive that I had to scroll to the bottom of the page to find the annotations. Then I had to scroll all the way back up to click on a different measure for other annotations. Bottom line--not an intuitive design for me. I would have expected the annotations to appear on the score, not in the lower left hand corner. ● Had to ask about where the audio annotations were although it seemed obvious after [the researcher] pointed it out. The pastelish color of the buttons made me think originally that they were "greyed out" or inaccessible, but I figured out that wasn't the case after accessing the audio annotations.
<p>Creating Own Data</p>	<ul style="list-style-type: none"> ❖ I had trouble replying and then couldn't really see the reply or set permissions. The replying was also in a very strange place. ❖ As a music theorist, I often apply Roman numeral or pc-set analysis to a score. The musical content that these labels refer to is very specific. The location of the annotations in the software is not specific enough. When you click on a measure, you can see all the annotations for that measure, but you cannot see what they refer to specifically. This is not a good tool for analytical annotation. ❖ Everything visible should be annotable. ❖ When I submitted a new description I got the loading sign but had to click on something else for it actually to load. I'm unclear as to why you can select multiple boxes at the same time. What does it mean to have them all

	<p>clicked? Also, should have mentioned this in the previous section, but I have to open the audio in a separate window for it to play...it gives me a message "there was a problem playing this audio" in the actual player on the site.</p>
<p>What aspects of the software did you particularly enjoy?</p>	<ul style="list-style-type: none"> ❖ Ability to toggle between score and other materials (manuscripts, recordings, etc.). ❖ Incredibly responsive. Liked the ability to see others' comments, as well as my own. ❖ I liked being able to see comments, but it was often tough to make work in a straightforward way. ❖ I enjoyed the potential of the software. I did not enjoy anything about its current form. ❖ The quality of the music typesetting was good, being able to select individual noteheads etc. with ctrl-click also very useful. Feels good to have ownership over annotation data even if Solid platform is arcane or not really explained in the app itself. ❖ I think this is something I would definitely use; seems like a great way to keep annotations organized.
<p>What aspects of the software did you not enjoy or find particularly frustrating?</p>	<ul style="list-style-type: none"> ❖ Annotations field and recordings player are a bit low on the page. ❖ Real-estate tradeoffs within the browser window and responsiveness...I found myself wanting the score to stay vertically on the page related to which things I was using on the comments/annotation side. ❖ Replying to the text, audio didn't work, and making my own annotations wasn't terribly clear. ❖ The software is not intuitive. Annotations are divorced from the content they refer to. You click "Reply" and then you have to scroll up to type your response. Annotations cannot be related to specific musical events. ❖ Zooming was quite slow and I found it frustrating that I was not able to identify comment authors even when mousing over the (i) icon. Comments slow to load after submission/committing to Solid. But tolerable.
<p>What aspects of the software need further work?</p>	<ul style="list-style-type: none"> ❖ Options for printing off or PDF exporting? ❖ Some small tweaks on user interface, to make it more clear where the users attention should shift once they hover over certain elements or click reply on an already existing annotation. ❖ The reply, the audio player, the responsiveness. ❖ All of it. ❖ Overall look and feel but also ergonomics: page turn buttons closer to page edges? ❖ Other than the little things I mentioned in the previous sections, I don't have any other comments.
<p>Are there any changes that you would need to see made before adopting a tool like this in your own</p>	<ul style="list-style-type: none"> ❖ It'd need to be faster and more reliable. ❖ Freehand annotations would be amazing! It would be nice to have references "back" to the digital score in the linked HTML pages as I felt we were quite far from the main interface by the time I was listening to the chosen audio excerpts. ❖ I wish there was a way to see the annotations on the score directly. i know that long annotations wouldn't necessarily fit, but at least the first few words would help me remember where I've made annotations and help

work flow?	give me the big picture. I'm a very big picture and also tactile person (the type who uses color markers all over scores), so in transitioning to a digital tool I would at least want some sort of overview function. I'd also like to see a highlighter features where you could mark up the score using various colors to highlight motives, themes, etc.
In its current state, what types of situations do you think software like this could be useful?	<ul style="list-style-type: none"> ❖ Archival research analyzing manuscripts and adjustments ❖ Exploratory discovery of existing musicology scholarship. ❖ It'd be good for doing musicological projects, but ideally there would be a local option as well, to be able to use without web access. ❖ Collaborative assignments for music students. ❖ Music analysis, analysis of score-based performances/recordings, could even maybe be useful somehow in corpus studies further on in its development??? That is, if there were some way of counting repeated annotations over multiple scores.
What types of situations do you think this software could be useful if the changes you mentioned above were implemented ?	<ul style="list-style-type: none"> ❖ Teaching analysis courses with annotated scores. ❖ Easier use. ❖ Teaching, research [sic]. ❖ It could be very useful for analysis courses. ❖ Critical score edition preparation, scholarly collaboration, interactive exhibition.

Table 2.1. Selected feedback from participants in the Music Scholars Mahler use-case study

2.2. Digital Score Edition: Early Vocal Music Use Case

2.2.1. Aim of the evaluation study

The aim of the follow up study was to collect user feedback on a set of two features as applied to the 16th Century vocal repertory developed and detailed as part of the **Deliverable 6.8 - Mid-term evaluation**¹⁷ report. These two features were a search query that allowed users to view any score linked through the TROMPA contributor environment and the ability to use F-TEMPO¹⁸ to perform partial match searching. We solicited general feedback on the text.

Our goals mirror those reported in Section 2.1.1. In addition to demonstrating the capability of the DSE to integrate sophisticated query features (using metadata to select scores for display and the musical content of a score for a similarity search in an external database) the follow up study also allowed participants to explore the flexibility of the DSE in working with other genres of music.

¹⁷ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

¹⁸ <http://f-tempo.org>

2.2.2. Participants

2.2.2.1 Recruitment strategies

Participants for the follow up study represented a subset of those used in the first study. All participants who participated in the Mahler Use Case Study received an email in Month 36 of the project inviting them to complete the follow up study.

2.2.2.2 Participant characteristics

The sample consisted of four music scholars from the Mahler Use case study. See Section 2.1.2.2 for a more detailed description.

2.2.3. Study protocol

Protocol for the Early Music Use Case was identical in its general form to that reported in Section 2.1.3 with minor exceptions. Instead of giving their general reactions to the software and its annotation features, participants were asked to use two features not represented in the Mahler use case: (i) selection of a score for display from a curated list of candidates within the CE; and (ii) the ability to send the music content of the displayed score as a query to an external search engine. In this proof-of-concept implementation, (i) is provided as a pilot TROMPA Contributor Environment Task (see **Deliverable 5.3-2 - TROMPA Processing Library**¹⁹) while (ii) sends extracted data (processed within the interface) as a query to the API of the F-TEMPO early-music search system. (The F-TEMPO search engine currently indexes approximately 500,000 page-images from printed music of the 16th and 17th centuries, returning as results a list of links to the images; in principle, if a musical work exists in F-TEMPO's index, this allows a user to identify an unknown work, or to see original printed examples, or pages from other derivative works using the same musical content.)

Notable differences from participants' reaction to the follow up designs are shown in Table 2.2.

2.2.4. Study evaluation outcomes

Data from the quantitative portion of the Early Music Case Study are shown in Figure 2.2. The following figure plots a smoothed distribution for each question reported. Discussion of salient and meaningful responses is provided below.

¹⁹ https://trompamusic.eu/deliverables/TR-D5.3-TROMPA_Processing_Library_v2.pdf

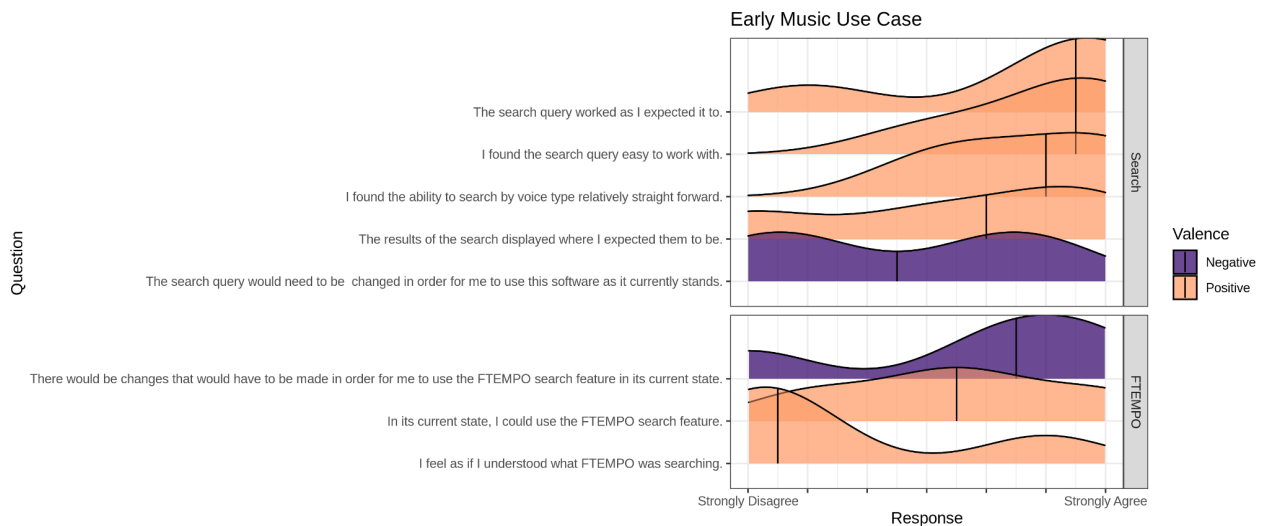


Figure 2.2. User Response Data: Early Music Use Case

In addition to Likert data collected as part of a survey, participants also had the opportunity to provide free text data to provide further context surrounding their experience with the DSE. We have included quotes that were particularly relevant to our research questions as detailed in the mid-term evaluation in Table 2.2. We then contextualize these in general discussion in Section 2.2.5.

Topic	Participant feedback
Search/Query Feature	<ul style="list-style-type: none"> ❖ 1) I found the presentation of choices slightly confusing (maybe because there was no heading to the page?) . It took me a moment to realize that there were three equivalent options to do the same thing, rather than three steps to follow in sequence. 2)The search feature searches while typing without hitting search--which was nice, if unexpected! 3) the list of all results is rather long to scroll through. ❖ The search could be improved with some auto-complete function (à la Google) and approximate searching (e.g. searching for "santus" instead of "sanctus" should provide similar results). I believe that there are out-of-the-box techniques/tools for this.
F-TEMPO	<ul style="list-style-type: none"> ❖ The labeling of the images did not make intuitive sense to me. There is considerable blank space between the name of the image and the image itself. I continued to click on items in the list and saw no images. I almost gave up until I accidentally scrolled down and saw them. ❖ 1)Before I searched, I had NO idea what "search this score on FTEMPO/search using FTEMPO" was going to do. I worked out that it was searching a partbook database (using information from the generated score?), presumably for concordances--but not all the results were

	<p>concordances? so, I still don't quite know 2) also, it's not clear that "search options" is going to take me to a list of voices (also, the blurb mentions searching for al voices, but it's not clear how one would do that.) 3) The search list is where expected, BUT the location of the images is dependent on one's browser window (which the modern score is not), so at first it seemed as if they were not there at all (in fact they were below the bottom of the modern score once I scrolled down.)</p> <ul style="list-style-type: none"> ❖ I interpret FTEMPO that it searches for similar melodic fragments. I have only checked the beginning of the Soprano voice. A more flexible searching function (e.g. specifying the pattern yourself) would be great! Or, at least, select motives (mouse click) in the Verovio score that are than automatically converted to a search pattern. Afterwards, it would be great to be able to make changes to the original search like "Disregard durations", "disregard pitches", "flexible durations (only relative)" etc. I was reminded of this project by Richard Freedman: http://digitalduchemin.org/search/ where such things are possible.
<p>What aspects of the software did you enjoy?</p>	<ul style="list-style-type: none"> ❖ Very easy to use ❖ Searching for items went smoothly and quickly! ❖ Rendering is relatively fast
<p>What aspects of the software did you not enjoy or find particularly frustrating?</p>	<ul style="list-style-type: none"> ❖ Wasn't clear what the FTempo was going to look for. Also having a hard time figuring out how to go back to previous page after searching FTempo ❖ Needs to be a little less laconic about what it is you are doing, and also the part book images didn't show up where I expected.
<p>What aspects of the software need further work?</p>	<ul style="list-style-type: none"> ❖ Needs to be possible to see the image and transcription side-by-side to be able to compare them. ❖ Being able to play the score is very helpful, especially since Verovio provides this functionality
<p>Are there any changes that you would need to see made before adopting a tool like this in your own work flow?</p>	<ul style="list-style-type: none"> ❖ I would like to be able to annotate the modern score as I examine the part books.
<p>In its current state, what types of situations do you think software like this could be useful?</p>	<ul style="list-style-type: none"> ❖ love the ability to compare score with manuscripts side by side to be able to see potential discrepancies ❖ Could be a useful "crib" for examining part books--a transcription is nearby; also useful for determining which pieces are contrafacta and which ones aren't. ❖ Corpus study on motifs; generally all kinds of similarities based on scores (historically, within/between composers,

	etc)
What types of situations do you think this software could be useful if the changes you mentioned above were implemented?	<ul style="list-style-type: none"> ❖ With better side-by-side view/annotation one could compare editions, ornaments, ficta, etc. which would be REALLY useful for performer/editors. ❖ I think that software like this has also great potential for teaching, especially with students who need to rely on GUIs

Table 2.2. Response data from Early Music Use Case

2.2.5 Joint study synthesis

The goal of both user studies was to use a guided questionnaire to assess the degree that we met the goals outlined in **Deliverable 6.8 - Mid-term evaluation**²⁰. In summary, these include the ability for participants to link annotations using a URL to various aspects of an MEI score, the ability to share annotations, the ability to select more than note and measure MEI elements, the ability to link annotations to specific directions in the score, integration of page flip and audio playing alignment, and the ability for individuals without formal music training to use the DSE. We first detail items in which our user test indicated we succeeded, then detail what goals we did not meet and reflect on why this is the case. We conclude with a brief discussion about future directions the DSE might take.

Several major successes can be reported from the Mahler Use case user testing. The first is the successful integration of features allowing users to link and annotate comments using various forms of data entry; these range from linking their own text, media, and URLs to the ability to reply to pre-existing annotation with the DSE. Second, central to the features specifically needed for the Mahler use case, the developer team enabled the interface to capture annotations inserted not only at the note and measure level, but also on tempo and directive markings in the score. This is particularly important for musicological work on composers such as Mahler, since a significant amount of scholarly discourse relates to how Mahler’s specific score indications are reflected in a particular conductor’s interpretation.

As shown in Figure 2.3, annotated performance directives are highlighted in blue. Annotations like these can then be linked either to text annotations or to images such as that of conductor Willem Mengleberg’s personal copy of the score, where he has scribbled out “Heiter” in the opening directive, probably on Mahler’s own advice (Figure 2.4). The DSE allows viewers to see where an annotation has been made, view it (in this case a highly significant performer’s handwritten annotation), then listen to various recorded performances.

²⁰ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf



Figure 2.3. Directive Annotations in Mahler



Figure 2.4. Mengleberg's Hand Annotated Score

We are also able to report several successes within the Early Music use case. One of our goals was to integrate into the DSE a metadata search feature for selecting scores for display that directly links with the TROMPA CE. This feature was explicitly tested in the Early Music Case Study with user response data shown in Table 2.2 In addition, we also succeeded in integrating external searches by similarity matching using the F-TEMPO tool (see section 2.2.3, above). This allows users to submit single voice-parts from a displayed score as queries to the F-TEMPO API for matching against a large index (c. 500,000 pages) of printed music from the 16c; links to images of the best matches are returned for viewing within the DSE. Participants here similarly expressed enthusiasm about the potential for this tool, specifically the tool's ability to link a score to original sources of the music.

Considering the joint successes of the two user tests, we are also able to report integrating with the Solid Pod technology that allows users to be able to have more security over how their data is stored. While this was implemented on a structural level, some participants noted that in addition to being able to annotate within the DSE, they would also like to be able to more easily download their own data in order to link the annotations made within the DSE to other environments. This is discussed further when talking about future directions, as this also aligns with the MELD framework goals.

Lastly, we were also successful in implementing features to allow users to navigate and zoom the score-display as planned in our mid-term evaluation. While we were able to implement these features on a general level, several participants raised issues with the current state of the zoom and

page feature in this prototype interface. For example, one remarked that they would like the zoom feature to have a finer level of control and the ability to reset to the initial level of focus. Another historical musicologist noted that in its current state it might be difficult to refer to specific pages within the score since its page-layout is not fixed: it is rendered to fit the screen at the chosen zoom level. These, and other, qualms with the current state of the software partly relate to some features not fully implemented in our prototype software, and partly to users' unfamiliarity with working in a purely digital domain.

We next detail some areas we were not able to meet as planned in our mid-term evaluation. The first major feature that we were not able to implement was to align and link an audio recording with the MEI score to allow automatic page turning during playback. In retrospect, the ability to implement this feature to the degree we had initially planned was limited more by re-prioritising our plans due to constraints of human resources and external pressures to the project (including the COVID pandemic) rather than any technical limitations; manual score/audio linking is extremely time-consuming, and developing and rigorous testing of an automatic process was simply not feasible. The developer team is aware how this feature can be implemented, and it serves as a clear point of departure for future work on the DSE.

Further, several participants raise specific issues with aspects of the user experience and user interface. Participants did not like that there was only one area to respond via text and that a text prompt did not appear where the subsequent annotation would be rendered. Further, depending on the resolution of the participant's screen, the query boxes and replies and annotations would not all be visible at once. The issue - items not appearing consistently where expected - was a recurring primary issue to be resolved for participants to adopt this tool in their musicological work and is an example of an important, yet in principle purely cosmetic, matter of interface design. While it does not yet fully fulfil our goal of an interface only requiring minimal formal musical training, it is clearly of high priority for further development.

Given these issues, the path for future work for the DSE is clear. The first and foremost changes to make to the current DSE prototype are with on-screen presentation and ease of use. For the software to be adopted widely, it needs to function more as individuals expect it to, which in turn demands more consideration of graphic design, which was considered beyond scope for the present prototype.

Almost all of the participants in the user study expressed a high degree of enthusiasm about the potential for what the DSE can do. Participants were keen to be able to browse more scores than just the default Mahler; the early music case offers this, but has so far been tested by many fewer users. Some participants saw further potential once a score and its annotations were accessible by a larger community of music scholars - as indeed the present interface allows, though such a community has not yet been built. In particular, when asked about alternative uses of the DSE several participants noted the potential for this to be used as a collaborative tool in pedagogical courses on music history and analysis.

In sum, early users of our prototype software showed a considerable amount of interest and excitement for the early prototype. In addition to having successfully carried out the majority of the basic functionality proposed in the mid-term evaluation, users of our software appeared to be able to see the applicability and future use cases of this use case. While there is certainly work to be done surrounding the users experience -- as is the case with any beta version software -- the next steps to pursue are clear.

3. Orchestras

Under the orchestras use case, three main user studies were run, and are reported in this chapter. First of all, following focus group sessions with the Delft Student Orchestra Krashna Musika, and members of various student orchestras across The Netherlands (reported in **Deliverable 6.8 - Mid-term evaluation**²¹), a third usability study was run, focusing on live campaign and transcription task design. This study informed the design of the prototype, reported in **Deliverable 6.4-2 - Working prototype for orchestras**²², that would be used for the subsequent user evaluations. As Deliverable 6.4-2 presented the application but no user evaluation outcomes, these outcomes are still reported in the current chapter, as Section 3.1.

Subsequently, in Section 3.2, we report on the evaluation outcomes following a series of user evaluations held with 19 youth orchestra members, over four evenings in March. One orchestra, the Almeers Youth Symphony Orchestra (AJSO), showed particular interest in our prototype, and volunteered to still be engaged more often in any evaluation studies. As a consequence, we still implemented final improvements to our prototype in the final month of TROMPA, and report on a final series of evaluations with the AJSO members in April in Section 3.3.

3.1. Third user study: live campaign and transcription task design

3.1.1. Aim of the evaluation study

This user study was performed in preparation of the final prototype deliverable for the orchestras use case, **D6.4 - Working prototype for Orchestras v2**²³. The goal was twofold:

- ❖ **Have users interacting with a live campaign**, that combined the refactored and updated components behind the Orchestras prototype (Campaign Manager, Crowd Task Manager, Scriptoria, and the CE as intermediate communication layer), for the ‘simpler’ tasks as part of the transcription procedure (clef, time signature and key signature recognition);
- ❖ **Improving task design for crowd transcription tasks**. In earlier iterations, transcription involved MEI code writing, which needed specialist knowledge, and made for a very user-unfriendly task. We therefore worked on more crowd-compatible transcription tasks.

As for transcription, breaking down this complex task into smaller, crowd-compatible tasks had led to a few possible task designs, as described below.

Option A: first encode generic notes and rests, followed by pitch correction and note duration tasks, or the other way around.

This would be a transcription process in three task steps; first, participants would encode the presence of notes or rests in sequence, as illustrated in Figure 3.1. Subsequently, the participants would either first input note durations (sequence rhythm), followed by pitch correction, or the other way around. For this, we wanted to investigate whether participants would feel it more natural to

²¹ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

²² https://trompamusic.eu/deliverables/TR-D6.4-Working_Prototype_for_Orchestras_v2.pdf


²³ https://trompamusic.eu/deliverables/TR-D6.4-Working_Prototype_for_Orchestras_v2.pdf

have a duration entering task based on a series of equal notes (Figure 3.2), or on a pitch contour (Figure 3.3).

Note Transcription: Position

The given segment to the left might contain notes and rests.
Use the buttons "Note" and "Rest" to represent the sequence of notes you see.

Example: note,note,note,note -OR- rest,note,note



Note

Rest

note,note,rest,note

Ready

Figure 3.1. Generic note-rest encoding mockup.





Note 1

●

○

◐

◑



Ready

Figure 3.2. Rhythm input on a series of equal notes.





Note 1

●

○

◐

◑



Ready

Figure 3.3. Rhythm input on pitch contour.

Note Transcription

The given segment to the left might contain notes and rests.
In the text area to the right, indicate the type of note (c,d,e,f,g,a,b) or rest(r) with the appropriate letter, while **separating by comma (,)**.

In case of an accidental, add "s" for #, "f" for ♭ or "n" for ♮ after the note letter, e.g. fs or ff or fn



Figure 3.4. Generic note-rest encoding mockup.

Option B: encode notes with pitch within the same task, followed by note duration correction.

In this design setup, the first task would ask for a participant to already encode pitched notes through text input (Figure 3.4), after which a rhythm input task similar to Figure 3.3 would be given.

A possible advantage of this setup would be that it is one task less, and that it may yield a quicker and more natural entry process for people who can reasonably read notes, where Option A may more artificially separate different task steps. At the same time, allowing for open text input may be risky from an input validation perspective, and reading speed may differ, depending on how familiar a user is with a given clef. In that sense, Option A may more universally be accessible.

3.1.2. Participants

3.1.2.1 Recruitment strategies

For this study, the regular list of orchestra contacts was approached by the RCO, representing 21 student orchestras across The Netherlands. As such, we had a potential reach of more than 1000 young musicians. We used general board contact information, unless we had a more direct contact into the orchestra from the previous user studies. The orchestras were offered 3 possible time slots for 3 different days. Ultimately, 7 candidates subscribed for the evening of February 16; out of these, there were 2 no-shows, such that we ultimately worked with 5 students.

3.1.2.2 Participant characteristics

As in the mid-term evaluation studies, we held a survey asking participants for their occupation and their level of music expertise through a selection of questions from the Goldsmiths Musical Sophistication Index (Gold-MSI). The compiled form of questions was:

- ❖ Please fill in your current occupation.
- ❖ I have had formal training in music theory for ___ years.
- ❖ I have had ___ years of formal training on a musical instrument (including voice) during my lifetime.
- ❖ I can play ___ musical instruments.

- ❖ The instrument I play best (including voice) is ____.
- ❖ I have experience designing User Interfaces.

All participants were students, connected to two student orchestras. All had extensive music-making experience, with more than 10 years of musical training. With two violin players, a clarinetist, a French horn player and a double bass player, the players also had experience with different clefs, in some case including experience with transposing instruments.

3.1.3. Study protocol

The study was conducted as a focus group discussion. First of all, informed consent was obtained and the musical background survey was held. After a round of introductions of the participants, a short introduction to the TROMPA project was given, after which participants were linked to a running campaign, on which they were asked to work silently for 10 minutes.

After this, a short break was held, followed by the Post-Study System Usability Questionnaire (PSSUQ) that also was used in previous usability studies under this use case. The PSSUQ consists of the following questions:

- ❖ Overall, I am satisfied with how easy it is to use this system.
- ❖ I was able to complete the tasks and scenarios quickly using this system.
- ❖ I felt comfortable using this system.
- ❖ It was easy to learn to use this system.
- ❖ The system gave error messages that clearly told me how to fix problems.
- ❖ Whenever I made a mistake using the system, I could recover easily and quickly.
- ❖ The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.
- ❖ It was easy to find the information I needed.
- ❖ The information was effective in helping me complete the tasks and scenarios.
- ❖ The organisation of information on the system screens was clear.
- ❖ The interface of this system was pleasant.
- ❖ I liked using the interface of this system.
- ❖ This system has all the functions and capabilities I expect it to have.
- ❖ Overall, I am satisfied with this system.

Following a discussion of the system, a presentation was held showing the different mockup options for the note transcription task. Participants were asked to comment on the designs, and suggest preferences on these.

Subsequently, to close the session, a discussion about campaign potential was held: would participants realistically see campaigns being run in the context of their associations, and if so, whom would they expect to participate?

As a token of gratitude for their time, as in previous studies, participants were offered the choice between a membership to Entrée, the RCO's youth audience association, or an RCO CD.

3.1.4. Study evaluation outcomes

The campaign ran successfully, and through the PSSUQ, the five participants indicated the system was easy to use. 100% of the participants voted 'Agree' to the statement "Overall, I am satisfied with how easy it is to use this system." As shown in Figure 3.5, the participants also indicated it was very simple to use the system. At the same time, this may have been steered by only the simplest tasks having been included in the campaign.

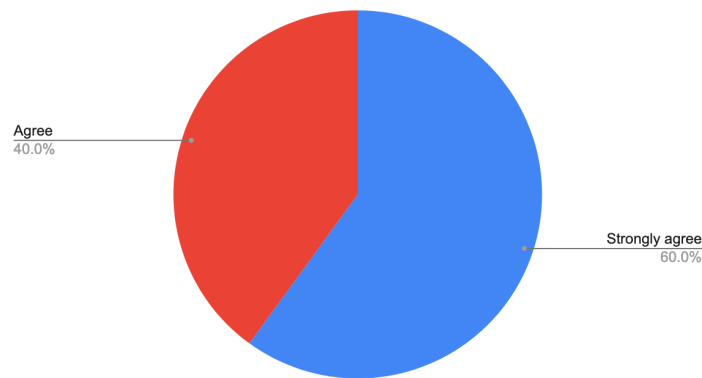


Figure 3.5. PSSUQ, February 16, responses to "It was simple to use this system."

In the feedback to the tasks, participants praised the simple task design, but did indicate they felt there was too much clicking on empty input (i.e. measures that do not contain clefs, time signatures, or key signatures). As clef, time signature and key signature changes do not solely occur at the start of a music system, but also may occur throughout a music page, we currently still require for each segmented measure to be checked. Participants suggested that in the future, one may add an extra task, similar to tiled image labeling captchas, in which several measures may be shown at once, and a participant would then check which of them contain clefs, time signatures, and key signatures. This feedback has not yet been prioritized for further development improvements though, as task management currently is strongly linked to individual segmented measures, but not to groups of measures.

Some concern was raised in the group on whether the shown tasks would not be *too* simple (and too repetitive) for people who can reasonably read music. One participant indicated they actually appreciated spending some effort interacting more deeply with 'traditional' music notation software, as it helped them to learn to use the software, and get into all the details of the music. Not in the context of an official user experiment, but when discussing the TROMPA context with an interested musicologist colleague (Mark Gotham), a similar observation had been made in the ScoresOfScores²⁴ project. This project has been associated with MuseScore's OpenScore initiative, and requested for the crowd to manually transcribe Lieder transcriptions. According to Dr Gotham, a motivation of people contributing to this initiative was to interact more deeply with music notation software. However, this is quite a different use case from the use case we have been studying in TROMPA, in which we explicitly seek to transform the transcription procedure into a hybrid, distributed and microtask-based procedure.

When discussing the transcription tasks, participants indeed preferred option A over option B, as that would allow for them to transcribe music in any clef (where in option B, the task easiness would

²⁴ <https://github.com/MarkGotham/ScoresOfScores>

indeed depend on their reading prowess). However, they were unsure about the insertion of generic notes before pitch or rhythm correction, and ultimately suggested to have the first task being a *generic rhythm transcription* task. In other words, they suggested for the first transcription task to consist of a constant-pitch transcription, in which notes and rests already would be at the appropriate duration lengths. We decided to take this design forward into the final prototype release, and the user studies to be conducted in March.

When asking for campaign potential, the group was hesitant, and actually indicated they were not sure if they could motivate others in their association to contribute. They especially were not sure about the engagement of non-musician fans in their circles who would join their concerts (e.g. parents, friends); in their opinion, these people would probably prefer to pay a ticket, rather than spending time on labor. This is in contrast with sounds we received in earlier, larger focus groups about the campaign setups, where different groups indicated they could see fellow players and audience members contributing. While this shows campaigns may not trivially be run and scaled, we believe that fatigue and demotivation induced by the ongoing COVID-19 crisis may have played an additional role in this feedback.

As one final, unexpected outcome of this session, one of the group members mentioned the Almeers Jeugd Symfonie Orkest (AJSO), a youth orchestra in Almere, that already has actively been playing from iPads and working with digital scores. As this orchestra is a youth orchestra and not a student orchestra, it was not in our contact list yet. Following this user study, we reached out to the AJSO orchestra's librarian, presented the TROMPA project, and got a very enthusiastic response back, with the librarian indicating high willingness to get people from the AJSO orchestra together for future user studies. As can be seen in the remainder of this chapter, AJSO would play a major role in the continuation of our user studies.

3.2. Prototype evaluation sessions

Following the feedback on our third user study, we still made several adjustments to our prototype. We kept the clef recognition (Figure 3.6), time signature recognition (Figure 3.7) and key signature recognition (Figure 3.8) tasks, while making the transcription phase consist of rhythm transcription (Figure 3.9) followed by pitch correction (Figure 3.10). For the final evaluation sessions in March, we wanted to test this full sequence of tasks, in efforts to complete a coherent music segment. For this, we focused on transcription of the first pages of the Beethoven Wind Sextet, op. 71. This music has the advantage of having multiple parts in different clefs and keys (due to transposing instruments), a time signature change on the very first page, but single-voiced parts throughout.

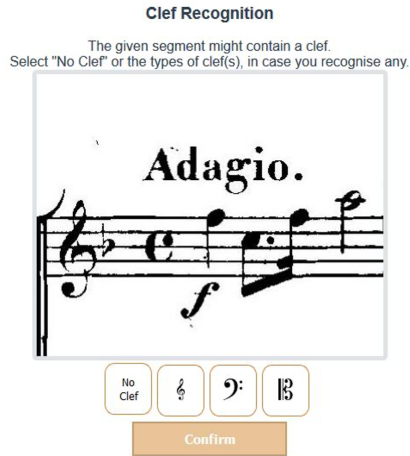


Figure 3.6. Clef recognition.

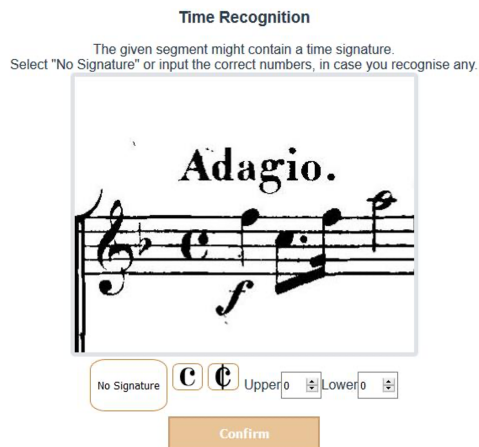


Figure 3.7. Time signature recognition.

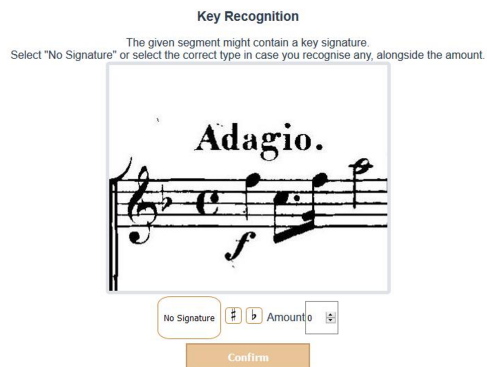


Figure 3.8. Key signature recognition.

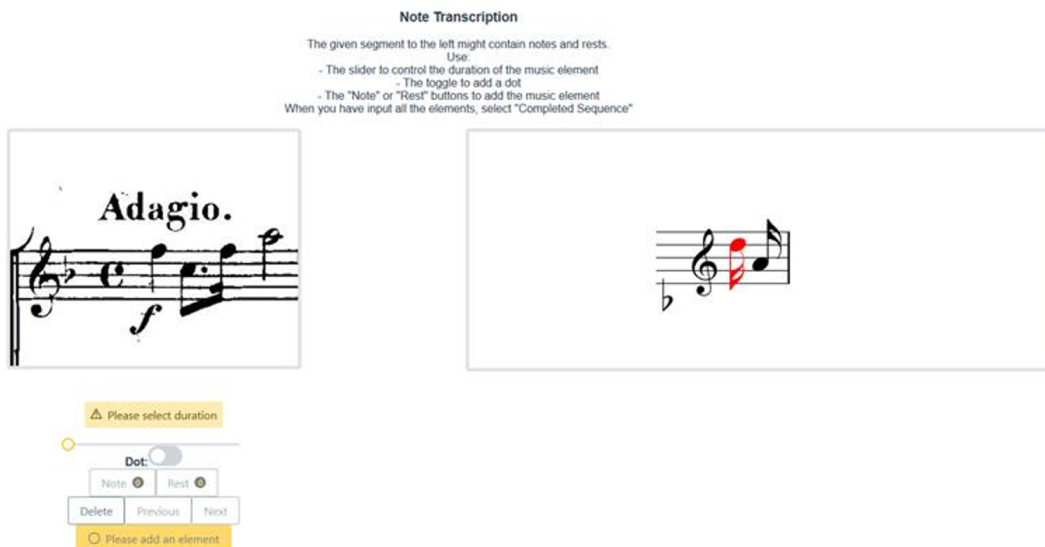


Figure 3.9. Rhythm transcription.

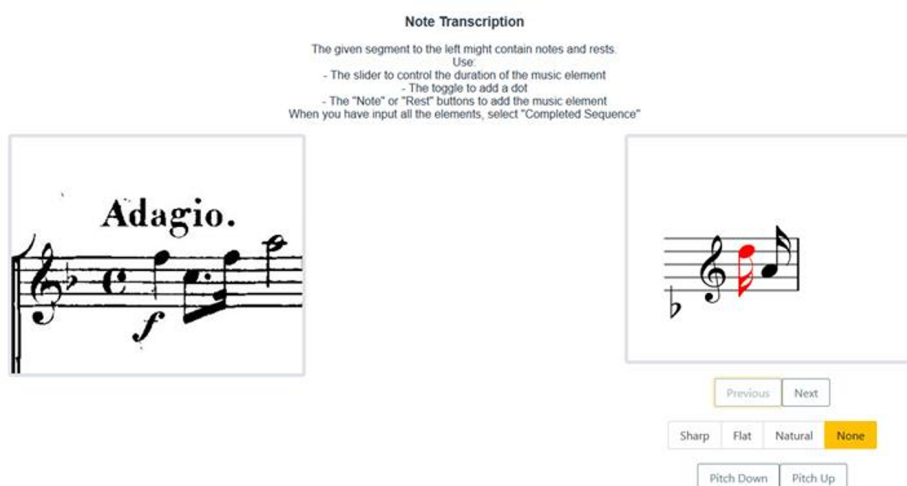


Figure 3.10. Pitch correction.

3.2.1. Aim of the evaluation study

For this series of evaluation studies, we had multiple aims:

- ❖ **Have users working on a live campaign with several phases in parallel**, towards concrete completions that can be rendered;
- ❖ **Gain understanding of the time needed towards task completion;**
- ❖ **Identifying remaining technical and usability issues.**

3.2.2. Participants

3.2.2.1 Recruitment strategies

Following our new contact with the AJSO orchestra, this orchestra recruited a group of five participants, who were available for a user study on March 16, 2021. Recruitment was initiated by

the orchestra's librarian. Being a youth orchestra, the orchestra's youngest members are minors (teenagers); for the recruitment, we explicitly requested to only include adult orchestra members (aged 18 or over), in order not to need extra human research ethics approvals, and any parental consent approvals.

Beyond the AJSO user group, we also wanted to run our evaluations with members of the student orchestras. However, we had some concern about recruitment success, given our difficulties recruiting a reasonably-sized group for the third user study in February. For the mid-term evaluation studies, it had not been hard to recruit 38 participants for 6 workshops. For the third user study in February, our recruitment strategies had remained the same, but it was much harder to get participants. Many former orchestra members (and even whole orchestras) who earlier indicated interest to be kept in the loop, did not react at all anymore. As mentioned in Section 3.1.2.1, we ultimately only managed getting 5 participants (and 2 unexpected no-shows) for a single workshop. Beyond TROMPA, such extremely low response rates have been observed more broadly in the current season (e.g. in MSc thesis project studies at TU Delft), and likely are due to demotivation induced by the COVID-19 crisis.

To still try maximizing the amount of participants, we tried to ease subscription procedures at the side of the participants. Where in the past, participants were asked to indicate availability on several possible time slots, after we as organizers would group them, currently, we offered a 'one-shot' subscription process, in which participants could directly book a preferred time slot. For the available time slots, we offered 10 possible evenings in March. In case only 1 participant was available on an evening, we contacted this participant to reschedule to an evening during which more people were already available.

Even with this setup, we still received a relatively low amount of responses. Ultimately, next to the AJSO study on March 16, we could only recruit for three evening workshops on March 24, 25 and 26, 2021, with 5, 7 and 2 participants, respectively.

3.2.2.2 Participant characteristics

Similarly to the previous studies, we ran a general music background survey, asking for the participant's occupation and musical background.

The AJSO session on March 16 involved 5 people, one being the orchestra's conductor, and one being the orchestra's librarian, with the other three members being students or young professionals. The librarian called in from an iPad, and indicated he could not read musical scores; we nonetheless invited him to join as a participant, and still try completing the tasks.

The session on March 24 involved 5 participants, coming from 4 Orchestras, with all participants being students or young professionals. The session on March 25 involved 7 participants from 3 orchestras, while the session on March 26 involved 2 participants from 1 (project-based) orchestra. 7 of these participants returned after earlier focus group participations, while the 7 other participants joined a user study for the first time. These new participants would be offered a similar token of appreciation as past participants.

In sum, the sessions on March 16, 24, 25 and 26 involved 19 participants. As shown in Figure 3.11, many of these players had extensive musical instrument training, and represented a considerable diversity of instruments (Figure 3.12).

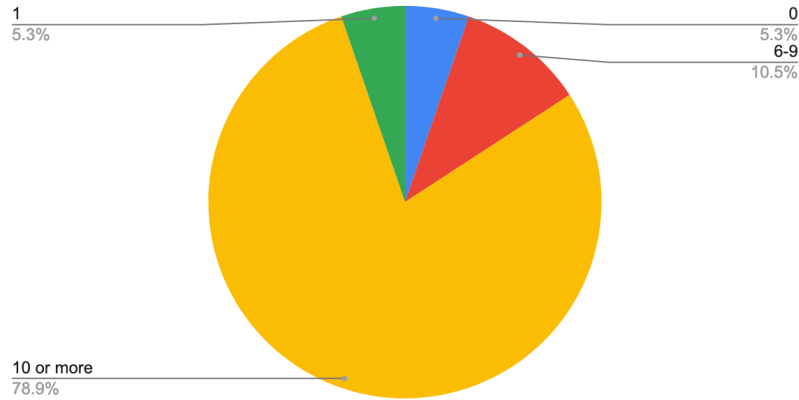


Figure 3.11. Responses to “I have had __ years of formal training on a musical instrument (including voice)” for the participants of the evaluation sessions in March.

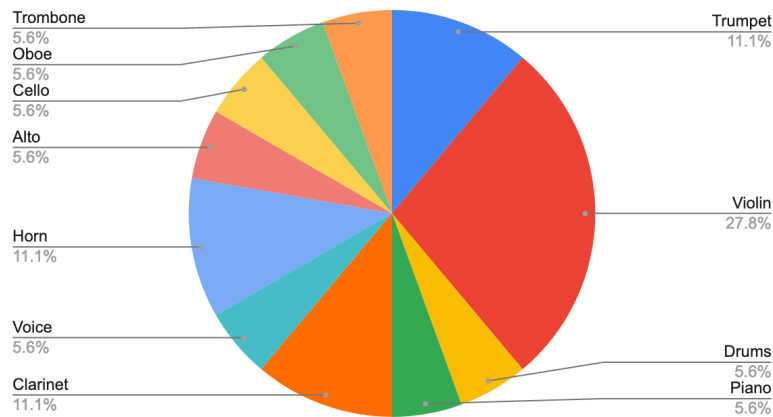


Figure 3.12. Responses to “I have had __ years of formal training on a musical instrument (including voice)” for the participants of the evaluation sessions in March.

3.2.3. Study protocol

The evaluation session combined unguided system interaction with group discussions. At the start of the session, informed consent was obtained and the musical background survey was held. After a round of introductions of the participants, a short introduction to the TROMPA project was given, after which participants were linked to a running campaign, on which they were asked to work independently for 60 minutes.

Considering the relatively small group sizes, dedicated campaigns were prepared for each of the groups, with an amount of tasks expected to be completable within an hour. To ensure that groups would see all the campaign task types, and not ‘hang’ too much in any task, we also set the amount of task completions required for an aggregation to 1. In other words, a submitted task for a given segment would immediately be considered completed, and not be sent to other participants, with the task only completing as soon as aggregation based on multiple inputs would pass.

On March 16, we had the 5 participants working on the first page of the Beethoven Sextet, which consisted of 120 segments (single-part segmented measures), that each had to be addressed in five campaign task types. On March 24, again having 5 participants, we offered a similarly configured campaign. On March 25, having 7 participants, we considered the slightly larger second page of the

Sextet, which consisted of 210 segments. On March 26, only having 2 participants, we only offered the first section of the first page of the Beethoven Sextet (7 bars), amounting to 42 segments.

While the participants worked on the tasks, in principle they would not be guided, and asked to work by themselves. However, the experimenters were available for questions and reports of technical issues, and while the participants were not encouraged to share system experiences, they were allowed to socially chat with each other.

After an hour of work, the experimenters would suggest for the work to be concluded, after which the PSSUQ was ran, that also was used in the previous studies. Subsequently, a walkthrough past each of the campaign task types was explicitly moderated by the experimenters, in which the participants were shown a screenshot, and asked to comment on their experiences with the particular task type.

Finally, participants were shown what happened ‘under the hood’, with the experimenters giving them a glance at their aggregated MEI code on Git, and rendering a MIDI of their produced result (which would sound strangely, as transposing instrument data was not yet encoded as part of the tasks, but as such gave a light-hearted end to the session).

3.2.4. Study evaluation outcomes

For each of the evaluation sessions, aggregated results are available on GitHub ([March 16](#), [March 24](#), [March 25](#), [March 26](#)). Cumulative time spent, considering the different task types, and based on commit data timings obtained from GitHub, is displayed in Figure 3.13. Average completion tasks per task type (in seconds), again based on available GitHub commit data, are given in Table 3.1.

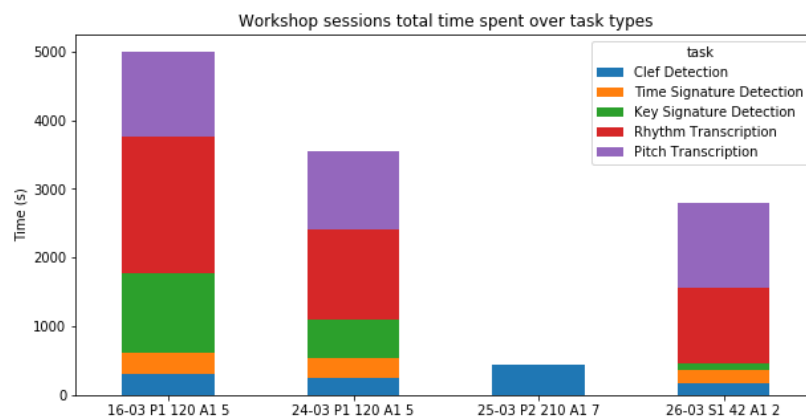


Figure 3.13. Cumulative time spent over task types, for the four different workshop sessions held in March. Each session is labeled with the session date, the annotated part of the Beethoven sextet (P1 = first page, P2 = second page, S1 = first section), the amount of available segments, the amount of tasks required before aggregation and completion, and the amount of participants.

Date	Clef detection	Time signature recognition	Key signature recognition	Rhythm transcription	Pitch transcription
16-03	12.5	13.3	48	82.9	51.5
24-03	10.7	11.8	23.5	54.8	47.1
25-03	14.7	Not available	Not available	Not available	Not available
26-03	8.4	8.9	4.6	52.8	58.3

Table 3.1. Average completion times (in seconds) for the different task types in the March evaluation sessions, based on GitHub commit data.

As can be seen from Figure 3.13, the session on March 16 lasted longer than an hour, also due to an unexpected system crash occurring during the key signature recognition task phase. With the timing data being inferred from available commits, the crash moment was not exactly logged, and adds to the completion time durations, also affecting the ‘average’ completion time for key signature recognition on March 16, as shown in Table 3.1. Furthermore, average completion times also are longer than the true time spent on a task, due to the relatively slow loading time of Verovio (typically, 5 seconds or more).

At the same time, the AJSO members were patient and reacted enthusiastically to the campaign, even when the system crash happened. When asked whether they would wish to stop after an hour, they voluntarily indicated they wanted to continue until completion. With the technical issues having been fixed in later sessions, the sessions of March 24, 25 and 26 did manage completing within the hour. Unfortunately, for March 25, due to a problem with the GitHub module, no commits were made after the key signature phase, causing us to not be able to indicate timings for that session. However, our task server still received the input, so a completed MEI file could still be made and shown.

As expected, with the more complex transcription tasks added, the system became more difficult to use. This both reflects in longer completion times for rhythm and note transcription, as well as in the PSSUQ responses, in which the system was not deemed as easy to use as in previous studies, that omitted the transcription task (see Figure 3.14 and 3.15). Furthermore, participants did not find all the functionality as self-explanatory anymore, as reflected in Figure 3.16.

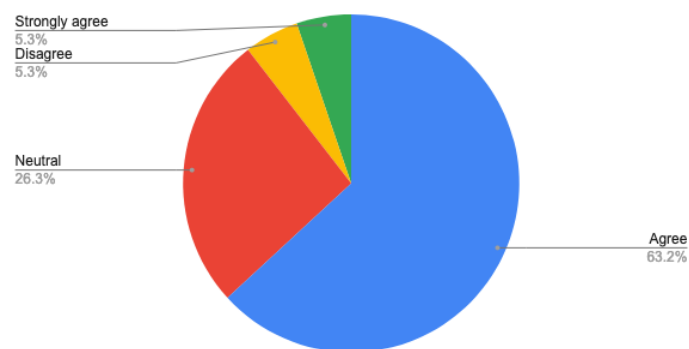


Figure 3.14. PSSUQ, March studies, responses to “Overall, I am satisfied with how easy it is to use this system.”

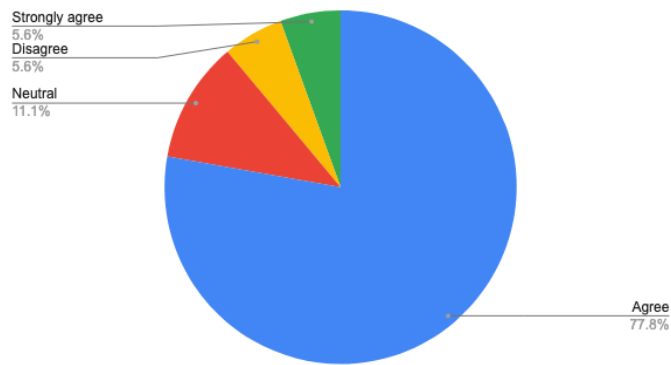


Figure 3.15. PSSUQ, March studies, responses to “It was simple to use this system.”

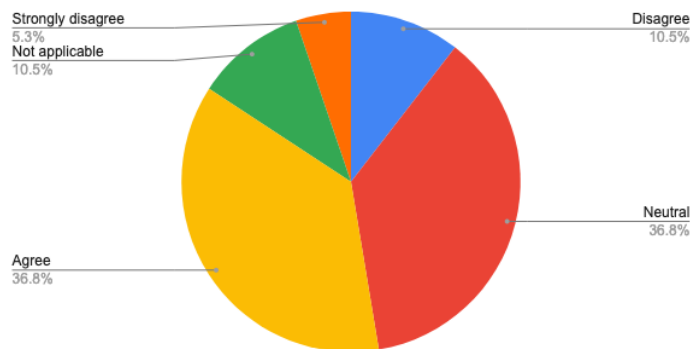


Figure 3.16. PSSUQ, March studies, responses to “The information provided with this system was clear.”

Discussing the participants’ reactions per task, the clef, time signature and key signature recognition tasks still were deemed to be relatively clear, although remarks were made about inconsistencies of task completion confirmation buttons across task types. Furthermore, for key signature recognition, some participants initially (wrongfully) assumed they had to count any accidentals that were available, not knowing the exact Dutch translation of the English concept mentioned in the interface. Next to this, participants made some suggestions on how to improve the interface to make it even user-friendly, e.g. by offering commonly occurring time signatures as clickable buttons.

The rhythm transcription task was received badly: participants found it difficult, cumbersome in terms of user interaction, and unclear in terms of goals. Further discussion revealed that here as well, offering common choices with clickable buttons may be easier, rather than using sliders. Furthermore, participants indicated that the absence of indicating beams between notes was annoying, as it caused many separate notes with separate flags, which would not match the given pictures, and take a lot of space.

The pitch transcription task was better understood, although participants indicated it required a lot of clicking, especially for bass clef parts (with default pitches being initiated in the treble clef). Participants suggested to include keyboard shortcuts, to reduce the amount of clicks.

For rhythm and pitch transcription, clef rendering also was not always correct, as in the given context, clefs would not yet be recognized as the ‘currently valid clef’, but still be considered as a potential clef change indicator.

Globally spoken, helper text was not always found as clear, and suggestions were made to either offer a tutorial, or give more visual hints at what would be expected of a user.

The student orchestra members were reasonably enthusiastic about the system, though not very clear on whether a more sophisticated system would be a system they would be willing to adopt. However, the AJSO reacted extremely enthusiastically, even despite a major system crash having occurred during their session. The orchestra members were so enthusiastic, that they voluntarily offered to still join further evaluation sessions, for which they also promised to try recruiting further members. Furthermore, the AJSO librarian (who could not read scores) indicated he was capable of conducting all the tasks, and also gave useful feedback on the experience on tablets.

3.3. Final iterations with AJSO

Considering the enthusiasm of the AJSO orchestra, we took advantage of their offer to still join more sessions. At the end of March and beginning of April, we still implemented several changes as suggested by participants. Most notably, we:

- ❖ improved helper text, replacing it by animated GIFs with example completion guidances (Figure 3.17);
- ❖ made interface elements more consistent, most notably the 'Confirm' button;
- ❖ revised all the task designs, with major overhauls to the rhythm and pitch transcription tasks;
- ❖ included beams in the rhythm transcription;
- ❖ implemented keyboard shortcuts on pitch transcription.

Screenshots of all improved task designs are given in Figures 3.18-3.22. Furthermore, halfway April, when noticing timing data was missing from the GitHub commits considered in Section 3.2 in preparation for this deliverable, we expanded and refined task logging, such that more detailed diagnostic information would be available, also on failing tasks and no commits because of empty input (e.g. no clef being present).

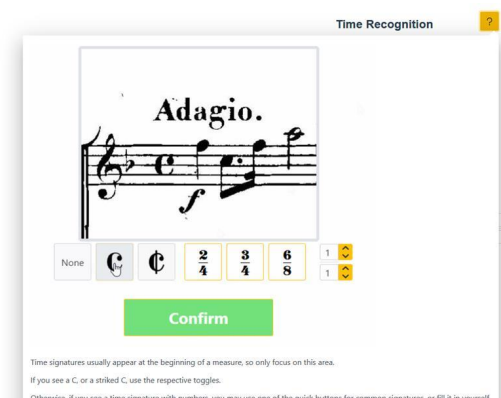


Figure 3.17. Improved help function: clicking '?' launches an animated GIF with visual examples.



Figure 3.18. Improved clef recognition. Compare with Figure 3.6, and note the improved 'help' and 'confirm' buttons.

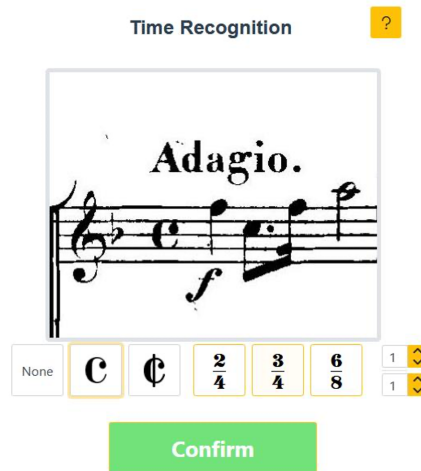


Figure 3.19. Improved time signature recognition. Compare with Figure 3.7, and note common time presets.

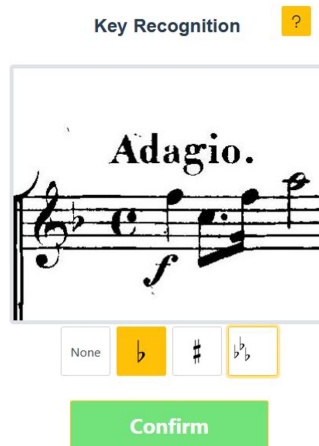


Figure 3.20. Improved key signature recognition. Compare with Figure 3.8, and note the more visual way of showing the amount of flats/sharps.

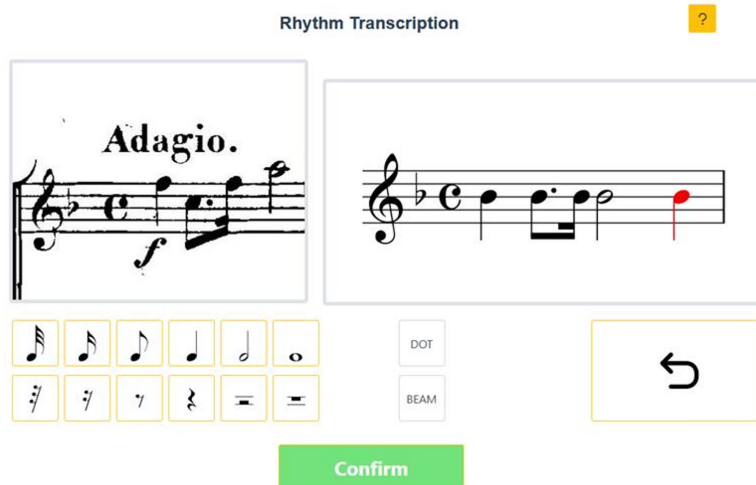


Figure 3.21. Improved rhythm transcription. Compare with Figure 3.9, and note the major interface design changes.

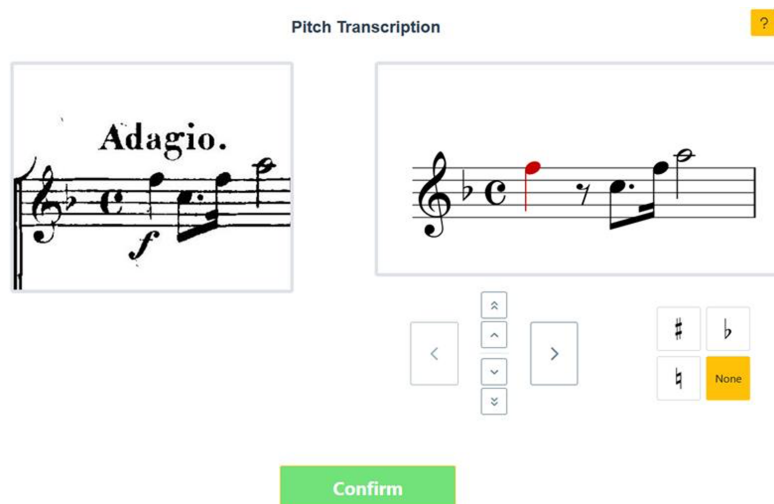


Figure 3.22. Improved pitch correction. Compare with Figure 3.10, and note the simpler button designs, including octave shortcuts.

3.3.1. Aim of the evaluation study

In two final evaluation sessions with members of the AJSO conducted in April 2021, we aimed to:

- ❖ **Understand whether suggested task improvements were effective**, especially with regards to efficiency of task completion;
- ❖ **Inform the feasibility of future scaled-up exploitation.**

3.3.2. Participants

3.3.2.1 Recruitment strategies

The AJSO members who had attended the session on March 16, worked on recruiting more orchestra members for sessions in April. On April 13, a small session (only considering the first section of the Beethoven sextet, which had a lot of note content in 7 bars) was conducted with 6 participants. After this, on April 20, a final session was conducted, in which the full first page of the Beethoven sextet was transcribed by 10 participants, including active aggregation, requiring for 2 task instances to be completed and successfully aggregated before considering a task as completed.

3.3.2.2 Participant characteristics

As with the previous study, participants were all adults, mostly with extensive music training backgrounds, with the exception of the AJSO librarian.

3.3.3. Study protocol

The study protocol was set up in similar fashion to the previous studies: after obtaining informed consent and conducting the music background survey, a short introduction to TROMPA was given, as each session included a few new members.

After this, participants would be invited to an hour of working independently but synchronously, with the experimenters being available for questions and technical issues, and with the option to socially chat. After completing the campaign work, participants filled in the PSSUQ, followed by a plenary discussion, walking through each of the tasks.

At the end of the session, we had a discussion with the participants on possible future adoption possibilities. As with previous sessions, new participants still received a token of gratitude from the RCO after having contributed to a session.

3.2.4. Study evaluation outcomes

Figure 3.23 shows the cumulative time spent on tasks, comparing the AJSO session on March 16 with the sessions on April 13 and April 20. It should be noted that the session of April 20 considered the exact same page that was digitized in the session of March 16, but with twice as many tasks, due to aggregation now being enabled. Average completion times are given in Table 3.2. GitHub results are publicly available ([April 13](#) and [April 20](#)).

As can be seen, completion was much faster for our improved system, with working sessions ending well within the hour. Average completion times on task types on April 13 still are higher than on April 20, as the first 7 bars of the Beethoven sextet (considered in the campaign on April 13) has a much higher note density than the second half of the page, which has many rests, and thus many tasks that can be very quickly completed.

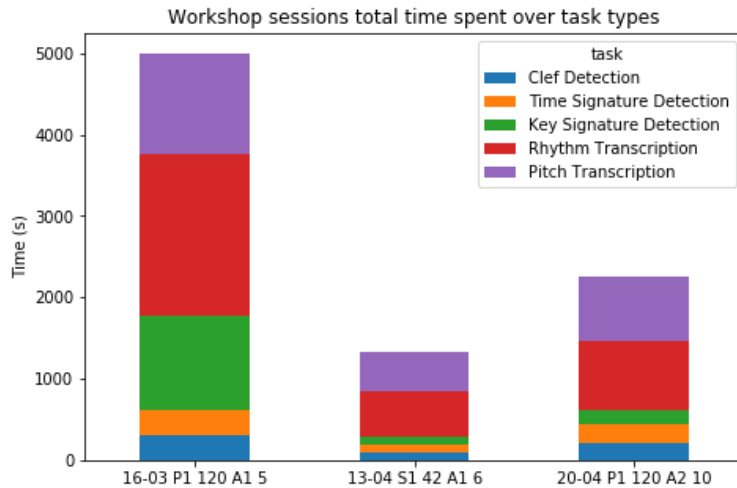


Figure 3.23. Cumulative time spent over task types, for the three evaluation sessions with AJSO.

Date	Clef detection	Time signature recognition	Key signature recognition	Rhythm transcription	Pitch transcription
16-03	12.5	13.3	48	82.9	51.5
13-04	12.3	15.3	13.9	80	69.3
20-04	8.8	9.4	7.6	35	33.1
20-04 (from system logs)	3.6	3.0	3.0	13.4	14.7

Table 3.2. Average completion times (in seconds) for the different task types in the three evaluation sessions with AJSO. All time calculations are based on GitHub commit data, unless indicated otherwise.

Taking advantage of our refined logging capacities, for the session of April 20, we also indicate more refined timing averages in Table 3.2, in which completion times for ‘empty’ tasks now also are explicitly counted. This shows major differences in the average completion time calculations, also indicating that ‘empty’ tasks (no clefs, only rests so no pitches, etc.) form a considerable part of a musical score. Based on the logs, we also give more insight in task completion distributions per task type (Figure 3.24), and how different task completions (completed, empty, failed aggregation) were served over time (Figure 3.25).

Both in Figures 3.24 and 3.25, we can clearly see the effect of many measures not having explicit clefs, time signatures, or key signatures, causing median completion times to be close to 0, and many tasks actually being ‘empty tasks’ without a commit. As discussed, currently, all measures are explicitly checked, as we cannot guarantee they are indeed free of clefs, time signatures or key signatures. However, in the future, an extra pre-processing task pre-selecting the measures with actual clef, time or key content (as already was suggested in the third user study, described in Section 3.1.4), would indeed help in reducing the amount of clicking downstream.

As for aggregation and possible error correction, it can be noted that aggregation failures mostly occur on the transcription tasks that are indeed deemed more complex. However, they still occur relatively rarely.

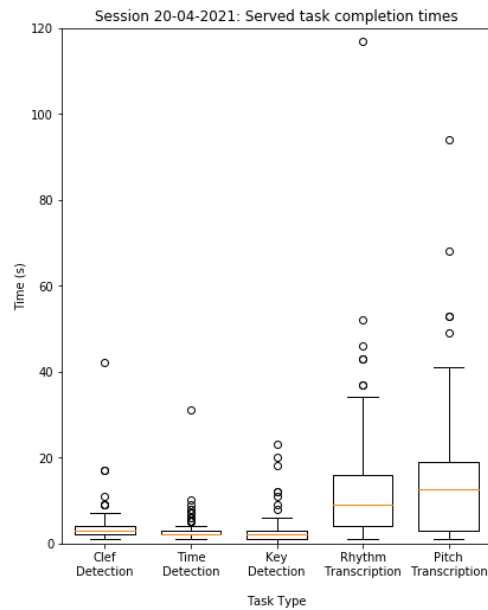


Figure 3.24. Task completion time distributions, based on system logs for the April 20 AJSO session.

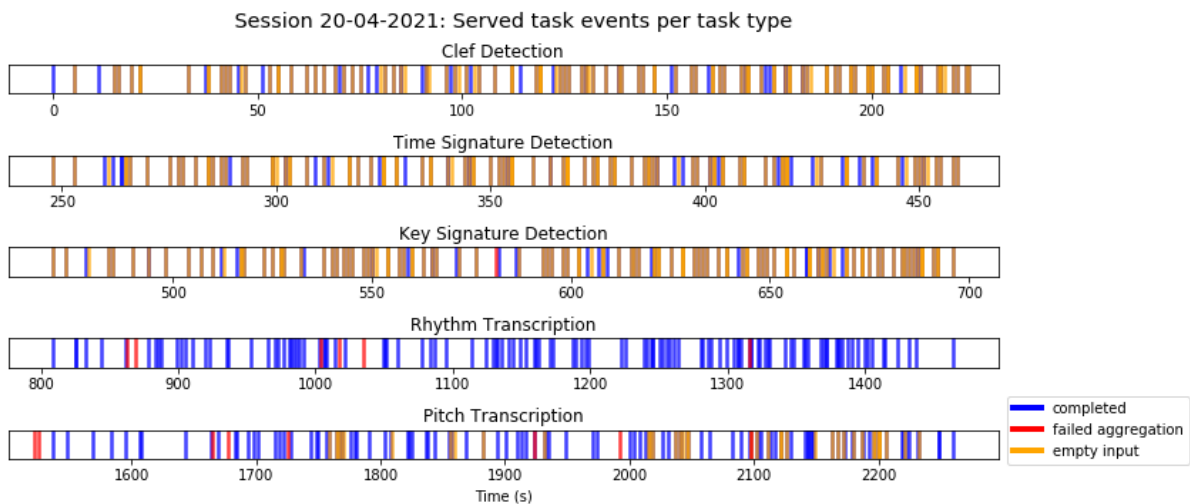


Figure 3.25. Task event distributions, based on system logs for the April 20 AJSO session.

Participants reacted very positively to the implemented improvements. PSSUQ outcomes (Figures 3.26, 3.27, 3.28) show much more positive verdicts on ease of use and availability of information, in comparison to the outcomes discussed in Section 3.2.4 (Figures 3.14, 3.15, 3.16).

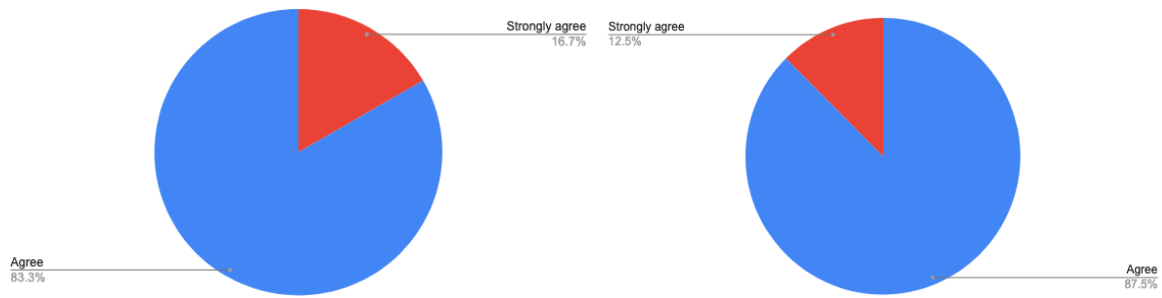


Figure 3.26. PSSUQ, AJSO studies April 13 (left) and 20 (right), responses to “Overall, I am satisfied with how easy it is to use this system.”

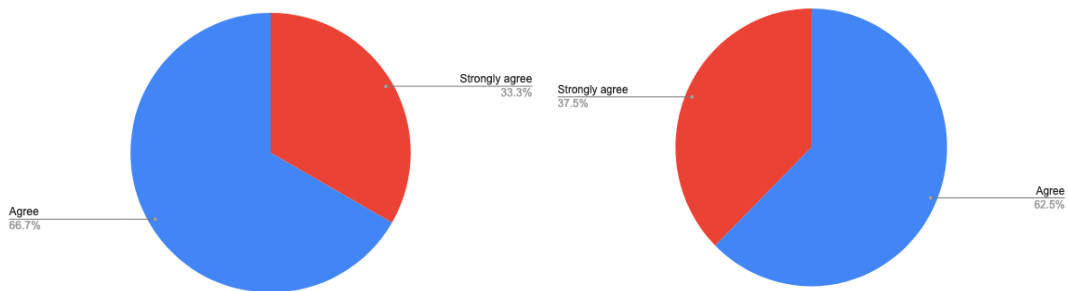


Figure 3.27. PSSUQ, AJSO studies April 13 (left) and 20 (right), responses to “It was simple to use this system.”

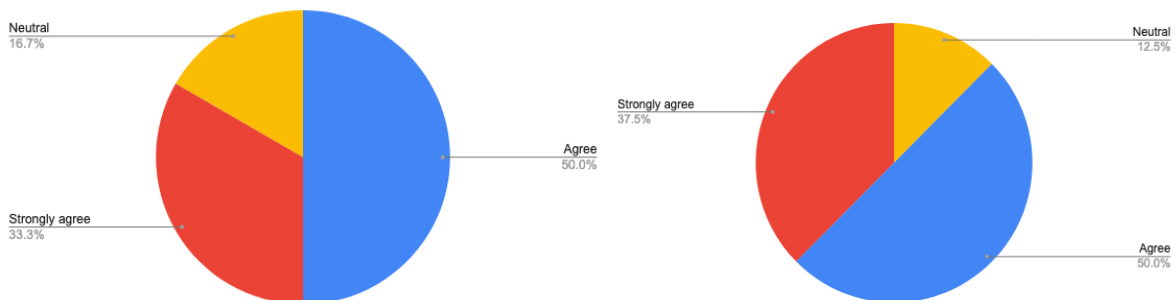


Figure 3.28. PSSUQ, AJSO studies April 13 (left) and 20 (right), responses to “The information provided with this system was clear.”

Stepping through the different task designs, the addition of beams and the newly added key shortcuts were highly appreciated. Rendering of clefs still has been problematic in some transcription tasks; no clear solution has been envisioned for this yet, as with the current knowledge that ‘a specific clef was seen within a measure’, we cannot yet tell whether this is a clef change indicator, or a ‘true’ clef counting at that moment. This also likely will require the addition of extra task types.

When finally discussing whether AJSO would potentially adopt campaign mechanisms and crowd transcription systems in the future, the orchestra indicated that more development would still be needed, for which they would not have bandwidth themselves. However, they would be very willing to still be involved as test users, in case the system would have a future, and they voiced hope that larger parties may ultimately pick up the concept.

Furthermore, as for the crowd working setups, the COVID-19 crisis had an unexpected advantage. In our evaluation sessions, we had chosen not to run ‘traditional’ crowd campaign setups,

in which participants would work asynchronously, and typically contribute short bursts of tasks, whenever they would be available and interested. Considering the difficulty in recruiting participants for the February and March studies, we were afraid that an asynchronous campaign would be easily abandoned, those never leading to coherent and complete results. Therefore, by letting participants work synchronously for an hour, we were guaranteed to see concrete outcomes based on several person-hours of work. AJSO indicated they very much liked this working setup: it gave them a remote, but social setup to jointly ‘get things done’, and indeed stick to a minimum level of commitment. In terms of future potential campaign designs and setups, this therefore may be a working format to also consider.

Finally, as an example of how the system improved thanks to the user tests, in Figure 3.29 we illustrate how display and output got much cleaner and readable, comparing several bars of results of the very first session with AJSO (March 16) with the final session.



Figure 3.29. Comparison of crowdsourced MEI input, AJSO March 16 (top, note no beams and a rest encoding bug) vs. April 20 (bottom, note better clef, rest and beam displays).

4. Instrument players

Following a small pilot study conducted using wireframe mockups in the first year of the project (reported in **Deliverable 6.1 - Final Mock-ups Testing**²⁵), two user studies were performed in scope of TROMPA’s instrument players use-case.

The first study, reported in Section 4.1, served the dual-purpose of i. gaining a richer understanding of the rehearsal habits, contexts, and information requirements of our target audience of musicians with advanced expertise in classical piano performance, and ii. obtaining user feedback on the initially implemented prototype developed in response to the outcomes of the pilot

²⁵ This deliverable is confidential to the consortium only.

study (reported in **Deliverable 6.5-1 - Working prototype for instrument players**²⁶), in order to inform further development. Due to restrictions imposed by the COVID-19 pandemic, this feedback was obtained in response to a demonstration of the interface presented remotely via video-conferencing.

The second study, reported in Section 4.2, employed participants from the first study and invited them to evaluate the final version of the implemented prototype, developed in response to their feedback from the first study. This evaluation took place interactively and in-person, adhering to all local requirements related to the COVID-19 pandemic.

4.1. Structured interview on digital piano rehearsal

The aims, recruitment strategies, protocol, and initial evaluation for this are described in **D6.8-Mid-term evaluation**²⁷. As only partial results were reportable at that stage, these details are reprised and outcomes are reported in full in this section.

4.1.1. Aim of the evaluation study

The aim of this study was two-fold:

- ❖ To obtain a more detailed understanding of pianist's information behaviours and requirements in the context of solo piano rehearsal, and to probe for opportunities around digital rehearsal tooling, particularly pertaining to the following topics (see 4.1.3 for further details):
 - Overall approach to rehearsing process when learning a new piece or perfecting one for performance
 - Rehearsal context
 - Rehearsal goals
 - Rehearsal activities
 - Interactions with students / teachers
- ❖ To obtain user feedback on a demonstration of the first implemented prototype of our performance companion

4.1.2. Participants

As the target audience for our prototype comprises musicians with advanced expertise in classical piano performance, we chose to recruit student pianists (piano majors) at the University of Music and Performing Arts Vienna (mdw).

4.1.2.1 Recruitment strategies

Participants were recruited through electronic postings on mailing lists, mdw social media accounts, and through physical posters placed around campus.

²⁶ https://trompamusic.eu/deliverables/TR-D6.5-Working_Prototype_for_Instrument_Players_v1.pdf

²⁷ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

4.1.2.2 Participant characteristics

Ten pianists participated in the study, evaluating the prototype as described in **Deliverable 6.5-1 - Working prototype for instrument players**²⁸. Participants tested exhibit a range of expertise, including five Bachelors or Masters students in piano performance and one with an MA in piano performance pursuing further postgraduate studies; one student with a Dr. artium (artistic doctorate) pursuing further specialised postgraduate study in chamber music; and three Bachelors or Masters students studying music pedagogy with a focus on piano. The participants spend between 8 and 40 hours on piano practice in a typical week (mean: 23.2, SD: 11.9), with a maximum of 16 to 55 hours (mean: 34.8, SD: 12.5). This diversity of experience corresponds to the scope of user audience envisioned for our application.

4.1.3. Study protocol

Due to the ongoing global pandemic, each session of this study was conducted remotely using the Zoom teleconferencing platform. Students participated in the sessions individually. Each session lasted approximately 1 hour, and involved experimenters David M. Weigl (CLARA developer, mdw) and Werner Goebel (pianist and performance scientist, mdw), alongside the participant. The interview was conducted primarily by D. M. Weigl, with additional contextualising and clarifying questions by W. Goebel applying piano performance domain knowledge.

At the beginning of each session, each participant was emailed three documents: an information sheet, a consent form, and a questionnaire (in English or German depending on participant preference). Participants were asked to read through the information sheet and consent form (see D1.3), and the terms of their participation were clarified. Upon consent, Zoom recordings were started to capture the session for transcription purposes. These recordings were not shared beyond the two researchers involved in this study.

Participants were guided through a series of questions on their rehearsal practice, focusing on the following subjects:

1. A general description of their *rehearsal strategies*, both when initially learning a piece, and when rehearsing a piece for performance.
2. The *context* of their rehearsal sessions – reflecting on how often, how long, when (time, weekday, context in terms of daily routine), and where (e.g., university, practice room, at home) they rehearse, and whether / how it makes a difference.
3. The *purpose* of rehearsal – what is being practiced? One or many pieces; whole pieces, or sections? Which repertoire, and how is this decided? Who guides the rehearsal – the pianist, or a teacher? Does it make a difference? Are specific objectives followed during rehearsal? Which (concrete examples)?
4. *Rehearsal activity* – what happens during rehearsal? Are rehearsals recorded? Are annotations made? Are digital tools used? If so, which? What properties must such tools have or not have, in order to be used? What can digital tools offer? What's currently missing?

The discussion on the above four points typically lasted about 30 minutes. At this stage, the CLARA prototype was demonstrated via Zoom screen-sharing, using some example rehearsal takes of a Clara Schumann piece recorded by W. Goebel for demonstration purposes. Participants were walked through each feature of the prototype, starting with a view of the rendered score; the selection and

²⁸ https://trompamusic.eu/deliverables/TR-D6.5-Working_Prototype_for_Instrument_Players_v1.pdf

playback of rehearsal recordings (demonstrating score-alignment through highlighting; dynamics and error visualisation, based on highlight colour; navigation of rehearsal recordings by clicking on score elements, or selecting larger structural segments from a drop-down menu; automatic and manual page turning; and finally, tempo curve visualisation, and navigation within and between rehearsal takes using tempo curves. Concepts around data ownership and sharing were briefly explained – that the data behind each rehearsal take is private by default, but that selected takes can be shared with specified others (e.g., teachers, colleagues) or made public, and that similar facilities are envisioned for score annotations.

Participants were given the opportunity to ask clarifying questions, and were then asked to reflect and provide honest feedback on the utility of the demonstrated prototype in light of the preceding discussion. Each session was then concluded with a final series of questions around pedagogical contexts:

5. *Pedagogical context* – could you envision using such a tool with your teacher? Could you envision yourself using such a tool when teaching your own students? What would be important in such uses? What properties would be required or need to be avoided?

Participants were then asked to return their filled in forms via email at their earliest convenience after the study is concluded. Upon receipt of their forms, they were sent a €20 voucher for Thalia, an Austrian highstreet and online bookshop chain, as a token of gratitude for their participation.

4.1.4. Study evaluation outcomes

Here we reprise and extend the initial outcomes reported in **D6.8 - Mid-term evaluation**²⁹. As per the study protocol (4.1.3), participants first responded on topics relating to their rehearsal practice, independently of our prototype which was only demonstrated to them after these responses were received. This was done to validate our assumptions and gather further user requirements, in order to guide the further development of the prototype.

- ❖ On *rehearsal strategy*, respondents differed in their reported initial approaches, with five indicating that they study the score of a piece (three incorporating listening to others' recordings, e.g., on YouTube) before beginning rehearsal, while the others began rehearsal renditions right away. Every participant mentioned the **annotation of fingerings** at a very early stage of rehearsal preparations, two participants mentioning deeper analytical processes (e.g., incorporating harmonic progressions) that take place before commencing with rehearsal practice. All but one participant mentioned practicing in a **slowed tempo** and building up speed to that anticipated for performance as rehearsals progress. In addition, one participant also mentioned playing slow pieces at an **increased tempo** and gradually slowing down to anticipated performance levels. Strategies for segmentation of the rehearsal process varied both across participants, and according to context for the same participant; though one reported a very strictly regimented split of rehearsal time into tightly focused 35-minute sessions, using the pomodoro technique³⁰. Two participants explicitly mentioned the **capture and study of (their own) performance recordings** at this stage, one particularly using this technique to identify missed directives (articulations, dynamics), while the other using it to sharpen focus during the recording process (using the prospect of putting the

²⁹ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

³⁰Cirillo, F. (2018). *The Pomodoro technique: The life-changing time-management system*. Random House. ISBN 9781524760700

recording online, e.g., on YouTube or Instagram, as an incentive). Two participants mentioned the similar benefits of performing rehearsal renditions in front of (one or several) others before a public performance, with one explicitly mentioning the effectiveness of “**simulated**” audiences (via video recording or Zoom call), which have become more relevant in the current pandemic situation.

- ❖ On *rehearsal context*, responses on typical daily **rehearsal hours** varied between 5 and 7, either in one session or split into two (morning and afternoon). Rehearsal **location** played a strong role for some participants and a weaker role with others – one participant asserting that practicing at home or at a dedicated rehearsal space at University makes little difference, others describing differences in rehearsal strategy (e.g., a greater need for focus, or a frustration with the limitations of personally-owned instruments, or concerns about neighbours and noise) when practicing at home, and one participant not owning their own piano and thus relying entirely on rehearsals in dedicated practice rooms. Each participant reported significant disruptions to their pre-pandemic routines in the current situation. Six participants reported access to **electronic (MIDI-capable) instruments** at home. The affordances of silent practice were outlined as advantageous by several participants (lack of neighbourhood disturbance; ability to validate and rehearse “muscle memory” and knowledge of the piece), while others regarded pianos that do not feel or sound like the grand pianos typically used during performance and during practice at University as inadequate for rehearsal purposes.
- ❖ On *purpose* of rehearsal sessions, responses differed, with one student reporting typical deep, focussed concentration on a single piece over the course of a 7 hour rehearsal session, while others typically rehearsed several pieces, often guided by the programming of upcoming concert or competition events. Two participants explicitly reported to never focus on just one piece, with one aiming to incorporate the learning of a new piece alongside rehearsal in every session. Each participant reported focussing on **specific sections** as well as on full rehearsal run-throughs, with one reporting that the former frequently turn into the latter (i.e., what was intended to be a rehearsal of a certain section just carries on to the end of the piece). Each participant reported some mixture of elements in terms of who decides **repertoire** for rehearsal – the pianist through personal choice, or an external factor (e.g., the teacher, an upcoming competition) – interestingly with varying effects on motivation, with one participant reporting motivation almost exclusively with self-chosen pieces, and another reporting the opposite. Two participants reported rehearsing with no explicit **goals** in mind (other than general progress: “*I rehearse what I’m not happy with*”; “*I want to play it better than I did yesterday*”). Others report specific goals, e.g.: “be loyal to the text” [incorporate specific metronome markings, articulations or dynamics as written in the score]; “*I want to play through and annotate fingerings on the first three score pages*”; “*play separate hands*”; “*fluidly perform these arpeggios*”; “*know precisely how to control your posture during a certain section*”; with one participant describing the pleasure of striking goals off on a checklist during rehearsal.
- ❖ On *rehearsal activity* – all participants **record** their playing at least occasionally; three using audio or video recordings routinely during rehearsal, and four others stating they “should” record themselves more often [because such recordings are deemed helpful]. One participant explicitly does not **revisit recordings** after immediate review, whereas another does so frequently, even consulting recordings from two years ago when a piece is rehearsed again after a long pause. The **utility of recordings** is appreciated by each participant.

Annotation of the score during rehearsal includes **fingerings** (all participants), **metronome indications** (one participant), **circling of notes** (six participants), **key indications** (one participant), and **structural annotations** (e.g., divisions, patterns in virtuosic passages; one participant). One participant mentions not writing anything beside fingerings, but incorporating annotations (e.g., **dynamic markings**) provided by the **teacher**. Another reports writing **short descriptive notes** outside of the **rehearsal context** (e.g., studying the score on a train ride). Another participant has developed an extensive system of symbols for a variety of purposes (e.g., accentuation; rubato), and mentioned the importance and meaning of colour in these annotations; and, the distracting and negatively perceived influence of having a teacher interfere with their own annotations. A further participant noted keeping more extensive **rehearsal logs** separately from the score inside a Google Docs document. Seven participants incorporate **digital scores** (displayed on an iPad) into their rehearsal practice. In each case, a bluetooth pedal is used to turn pages. Of the remaining participants, two would be interested in switching to digital scores but cannot do so at present for budgetary reasons, whereas the other explicitly prefers reading from paper. For users of digital scores, additional tooling includes the **forScore app** (score viewing and annotating app mentioned by two participants), and the **Henle app** (score subscription service). Six participants also indicate the use of **IMSLP** for score acquisition. In terms of properties required for a digital tool to be useful, **focus** and **ease of use** (lack of distractions, mentioned by all participants), **reliability** (mentioned by three), **performance speed** (mentioned by two) were seen as most important, particularly in a concert performance context. **Affordability** was mentioned by one participant.

At this stage of the interview, participants were given a demonstration of the CLARA prototype using some example rehearsal practice data (three recorded rehearsal attempts of Clara Schumann's *Romanze ohne Opuszahl*) generated by the experimenters for this purpose. The feedback this demonstration prompted can be summarised as follows:

- ❖ The ability to **revisit rehearsal recordings** and to **navigate** these through interaction with the score was universally seen as useful, as was the ability to **visualise performance errors**. **Automated page turning** was explicitly deemed useful by three participants, though one requested these to happen earlier than as in the demo shown to them (e.g., half a measure before the end of page, possibly configurable). Feedback on **tempo** and **dynamics visualisation** was mixed: three participants saw tempo visualisation as having limited (three participants) or no (one participant) utility; whereas six appreciated its usefulness, with one stating *“the feature that appeals most is the tempo”* (for checking variability and consistency with objectivity, and comparing to previous rehearsal renditions). Similarly, **dynamics visualisation** was seen as less helpful by three participants (*“should be audible, not visualised”*), but potentially very useful by four others (*“as a sanity check to your own perception”*; *“checking that every voice in a fugue is audible”*). One participant specifically proposed **aggregate dynamics measures** as a useful means of **performance error detection**, e.g., to verify that a crescendo specified in the score was reflected in performance – though outlined the need for the use of good editions to provide the necessary references. Two further participants outlined the potential usefulness of supporting explorations of tempo or dynamics in more granular terms, according to voicing or staff: *“What is the balance like in terms of the chords?”*. In broad terms, three participants saw utility for their personal rehearsal practice and could see themselves incorporating such a tool into their rehearsals;

of the remaining two, one participant (the most advanced pianist) could see utility of such a tool at earlier points in her career; whereas one saw utility only in the automated page turning, but otherwise stating that “*I don’t see how this would help me develop my muscle memory ... I don’t see how it would help me in practice*”.

Finally, participants were asked to reflect on the use of “such a tool” in a pedagogical context, both as a student and as a teacher.

- ❖ On *pedagogical context* responses were mixed. Five participants expressed concerns that teachers would not want to use such a tool due to the **additional time and effort involved** (also echoing comments during the user study pilot sessions in the first year of the project), whereas five saw potential for its use, with one reporting on experience with a digital platform that matches competition participants with experienced reviewers / judges as an example use case where such a tool would be particularly well placed. Participants were more open on the prospect of using the tool *as a teacher* with students, with three participants with younger pupils (teen-agers / young adults) responding particularly positively to this idea, one participant drawing out the need for new solutions given the current pandemic situation in this context.

4.2. Interactive evaluation of the prototype

4.2.1. Aim of the evaluation study

This study aims to evaluate both the additional developments undertaken in response to feedback obtained through the study reported above (Section 4.1), and the rehearsal tool as a whole at final state of development within the TROMPA project. Concrete improvements in response to feedback to the original study are detailed in **Deliverable 6.5-2 - Working prototype for instrument players**³¹ (Section 4.3), and include:

- ❖ Significant improvements to the speed of both initial load and page turns, intended to meet expectations on **reliability** and **performance** of digital tooling
- ❖ Creation of new dedicated visualisation modes for **performance errors** and **dynamics**, in addition to the indications through note colouring, meeting suggestions on the usefulness of surfacing this information more clearly by several participants.
- ❖ The creation of a further new visualisation mode allowing specific **performance instants** (e.g., chord soundings) to be illuminated in greater detail, zooming in to show relative timing and dynamics of individual notes within a chord.
- ❖ The ability to **show** and **hide** individual feature visualisations (as well as switching them off entirely), to meet the diversity of expectations and address the worries about distraction and complex interfaces.
- ❖ The implementation of basic **score annotation** functionalities, seen as useful to at least a certain degree by all participants.

³¹ https://trompamusic.eu/deliverables/TR-D6.5-Working_Prototype_for_Instrument_Players_v2.pdf

4.2.2. Participants

4.2.2.1 Recruitment strategies

As part of our aim was to evaluate features developed in response to feedback obtained in the first study, we chose to invite the same ten participants (all of whom had explicitly indicated a willingness to be contacted again for further research participation during the first study). We received six positive responses.

4.2.2.2 Participant characteristics

See 4.1.2.2. Six participants from this previous study participated in the interactive evaluation of the prototype.

4.2.3. Study protocol

As in the previous study, students participated in these sessions individually, alongside experimenters D. M. Weigl and W. Goebel. Due to restrictions imposed by the ongoing COVID-19 pandemic, the study was performed in hybrid mode, with one quarantined participant taking part remotely (via a Zoom conference call with screen-sharing), and five taking part in person, adhering to all local hygiene requirements at the time: the participants and experimenters obtained negative COVID-19-antigen tests prior to each session, and they maintained a 2-meter minimum distance and wore FFP2 masks throughout each session, which took place in a well-ventilated room.

Participants were given an information sheet and asked to fill out a consent form after being given the opportunity to ask for clarifications at the beginning of each session. In-person participants were then told they would be asked to engage in piano rehearsal using the latest version of the rehearsal tool demonstrated in the prior study, and asked to choose a piece for rehearsal from those available via TROMPA's repository of music encodings.³² These participants were then shown the prototype in "page view" mode, displaying a full page of their chosen digital score, on a computer screen placed on top of a digital piano (Yamaha Clavinova CLP-470). The piano was connected to the computer via a MIDI / USB cable. The participants were then asked to start an initial play-through. On completion of their rehearsal, the alignment process was briefly described, and the prototype was put into "feature analysis" mode to demonstrate its outcomes. The student participating remotely unfortunately did not have access to a digital piano at home, and so experienced the prototype using example rehearsal data.

Participants were then walked through the various features of the prototype in feature analysis mode: score-aligned playback with automated page-turning; navigation using player, score, or feature analysis curves; the tempo curve, dynamics summary, and error displays; the more detailed dynamics views, alongside the ability to show or hide individual analysis panes; the detailed instant view showing exact timing and dynamics of all notes sounded within a given chord; and the opportunity to generate simple score annotations, and to delete them. Participants were given the opportunity to experiment with each of these features, ask clarifying questions, and record further rehearsal attempts for comparison. Each participant recorded at least three substantial attempts (playthroughs or fragments), and the interaction with the prototype typically lasted around 30-40 minutes. Finally, participants were asked to fill out an evaluation form. This included a general evaluation section, involving rating (on a scale of 1 – 7) and commenting on the following aspects:

³² <https://github.com/trompamusic-encodings>

- ❖ **Usefulness** of the tool (“Not useful” – “Very useful”)
- ❖ **Usability** of the tool (“Very difficult to use” – “Very easy to use”)
- ❖ **Accuracy** of the information available (“Very inaccurate” – “Very accurate”)
- ❖ **Performance** (speed) of the software (“Very slow / unresponsive” – “Very fast / responsive”)

Further, the evaluation included a component evaluation section, involving commenting on the following components (“What did you like and/or dislike? What’s missing? Any other comments?”)

- ❖ Digital score displayed in the tool
- ❖ Support for score annotation available in the tool
- ❖ Tempo curve display
- ❖ Dynamics analysis display
- ❖ Error (inserted / deleted note) display

At the end of the evaluation, participants were asked to rate how **likely** they were to use the tool in their own future rehearsal practice (on a scale of 1 – 7, “Very unlikely – Very likely”), and given the opportunity to write down any further comments. Finally, participants completed a short questionnaire on their academic music background and rehearsal habits. After completing the study, participants were sent a €20 voucher for Thalia, an Austrian highstreet and online bookshop chain, as a token of gratitude for their participation.

4.2.4. Study evaluation outcomes

4.2.4.1 Rating responses

Overall, user responses on the **usefulness** of the tool ranged from 5 – 7 (median: 5). **Usability** was responses ranged from 4 – 7 (median: 6). The tool’s **accuracy** received ratings from 4 – 7 (median: 5.5), and **performance** (defined as “performance (speed) of the software”) ratings ranged from 4 – 7 (median: 5.5). Finally, users indicated the **likelihood** of themselves using the tool in their own rehearsal practice between 4 – 7 (median: 5).

4.2.4.2 Written responses

Participants were largely positive about the tool and its applicability and usefulness in the piano rehearsal context: *“I think it will be very useful for professionals and amateurs to analyse their own performance”; “certainly it can be a very helpful tool for musicians, both professional and non-professional”.*

While the tool’s overall functionality was praised by all respondents, the enthusiasm of several participants was moderated by limitations of the interface: *“the idea is very interesting and I think it will be extremely useful for many of us. Generally, I just think that it needs to be as easy as possible to use and fast”; “I think this program could be really useful ... the interface of the program could be more user friendly”.* However, this was not a universal view: *“Intuitive and easy”; “It seems very easy and fast”.* Several participants offered concrete proposals for improving the interface usability, including: increasing the font size of textual labels, or replacing them with icons; softening the colouring of the paging bar (currently a prominently placed bright orange element that was deemed distracting); making the note-highlighting during playback optional; including “help” buttons to explain the different features.

The accuracy of the presented information was largely accepted, but small inconsistencies were noted: *“It was very accurate almost all of the time, with a very few exceptions”; “in this moment many small notes can’t be shown. But most of the text was very correct”.*

The quality of the rendered digital score was praised by all respondents: *“Looks very clean”; “Looks great”; “Score is very clear and it’s easy to play and to work with”; “no visible differences from a paper score ... I found no problem performing from it and it was visually very easy to see all the notes, dynamics, tempo indications, etc.”*. Score annotation functionality was satisfactory to some: *“Annotation works very fine”; “It worked great and it can be very useful”; “It’s enough everything it has already”*. Others wished for further features, e.g. *“I would expect more (dots, binds, lines, pedal marking) ... [and] have multiple colours”; “Still quite simplified, however, it is very difficult to make it more complete without turning it into an excessively restrictive tool”*. On the latter point, this respondent along with another participant expressed a desire for support for free-form annotation (e.g., using a bluetooth pencil tool).

The utility of the tempo curve display was universally praised: *“Very precise and clear”; “Accurate”; “By far the most interesting and useful tool”; “Very useful especially when comparing different performances”*. However, it was deemed unnecessarily complicated by one participant: *“As an idea is great. I’m not so sure how easy it is to read and understand it”*. This participant proposed the removal of the “dot” indicators marking score positions, and proposed showing only the most recent performance by default, as a means for simplification.

The tools for dynamics analysis were largely praised: *“I believe it is very helpful as a second pair of ears”; “it worked very well ... a good tool when comparing different performances”; “Dynamics analysis works perfect. Very useful to work with phrases. Also to work on balance between hands and voices.”* However, one participant noted limitations of MIDI vs audio analysis when it comes to the accuracy of dynamics information in particular: *“unfortunately, the technology of pressure detection is still not even comparable to the result we’d get on an acoustic piano, so the accuracy of dynamics is not the best for high level”*.

The tools for error visualisation were also praised by most participants: *“Very useful”; “Great! Very helpful!”; “Inserted or omitted notes are detected very well”*. However, several participants noted small inaccuracies with the tool: *“Not 100% accurate yet, but very reasonable”; “A few times it showed that some notes were missed when they were played but other than that, it was very accurate”; “[Problems] with notes which are free in notation (trills, cadence)”*.

Finally, estimates of whether the participants themselves would be likely to use the tool in future ranged from neutral to enthusiastic. Several anticipated using the tool particularly in context of more complex pieces: *“I think this program could be really useful for very complicated piano pieces like Ravel ‘Scarbo’”; “It would definitely be very interesting to explore the software more, especially performing more difficult pieces with a more complicated structure”*. Others anticipated uses in pedagogical contexts: *“As a piano teacher to young non-professional musicians, I think it could be even more useful to explore and work with the full capacity of this really impressive program”*.

5. Choir singers

For the final evaluation of the Pilot, we established a collaboration with Cantoría. Cantoría (cantoriamusic.com) is a vocal quartet specialized in the performance of vocal polyphony from the Iberian Golden Age repertoire. The freshness, the youthness and closeness have become the distinctive traits of this ensemble, that begins to build a notable national and international career.

They have a project, *More Hispano*, where they organize participatory concerts. For instance more than 200 singers from 12 countries joined their voices with Cantoría in Christmas 2020.



Figure 5.1. Cantoría ensemble

In this collaboration, the TROMPA project joined Cantoría and the Escola Superior de Música de Catalunya (ESMUC, associated partner of TROMPA), to co-organize a participatory concert (face-to-face and virtual) using the Choir Singing Pilot (CSP), that we called "Cantamus" for a more appealing and easy to remember name.

The concert and related educational activities hosted by ESMUC allowed, on the one hand, to disseminate the musical repertoire of the Iberian Renaissance, of which Cantoría are experts and for which they have created the *More Hispano* project; and on the other, offering the possibility of participating in the concert, learning some pieces of the repertoire. Participants were able to learn these pieces using the Choir Singing Pilot. Prior to the participatory concert, we organized a pilot evaluation and a series of activities related to the concert.

For the preparation of this project, we carried out a set of recordings in a professional studio with Cantoría ensemble, and from them we implemented two of the identified functionalities in previous evaluations:

- ❖ **Instrumental track:** it is common that choir repertoire contains an instrumental accompaniment by orchestra or piano. This has been extensively requested in previous user studies. We implemented a new audio track to the CSP, allowing choirs to upload an instrumental track, in the form of an audio MP3/WAV file. This instrumental track is perfectly synchronised in time with the MusicXML score. For this project, we used the organ as an instrumental accompaniment.
- ❖ **Practice tracks** (expressive performances with varying tempi): in addition to the functionality of synthesising scores with artificial voices, some choirs requested the possibility to upload already available practice tracks³³. We implemented this new functionality with Cantoría tracks, in order to upload one audio file per part. We carried out manual score alignment with a time-varying tempo, as we find in real expressive performances.

³³ Practice tracks are audio recordings for each part (e.g. four tracks in the case of SATB scores) that choir conductors share to choir members. These tracks can typically be a singing recording or also a piano reference recording with the melody.

The target Cantoría repertoire for our evaluation consisted on three pieces:

- ❖ *Oy comamos y bebamos*, by Juan del Encina. It is a profane song, of popular origin, sung between the 15th and 18th centuries. It was sung by the so-called ‘villanos’ (from the word ‘villa’, town). Later, it began to be sung in the temples and it was associated with Christmas. These songs with profane lyrics were forbidden and Christmas carols with religious lyrics about the Birth of Jesus were sung instead.
- ❖ *Sus, sus, sus*, by Bartomeu Cáceres. It is a fragment of the composition "La Trulla" known as ‘ensalada’, a polyphonic musical genre of the time, which mixes, in the same piece, different musical styles, languages, textures and other elements of music.
- ❖ *Teresica Hermana*, by Mateo Flecha. The author is the most prolific composer in the ‘ensalada’ format. The composition is profane and it is integrated into the ‘Cancionero de Uppsala’, a book containing Spanish ‘villancicos’ (Christmas carols) from the Renaissance period.

However, due to time constraints, the community rehearsal was targeting the first two pieces. In this study we also addressed some technical needs that choirs expressed in the current COVID-19 situation, and tried to address them with the Choir Singing Pilot:

- ❖ The possibility to organize **virtual choir rehearsals via VC**: in order to address that, choir conductors were using the Choir Singing Pilot to reproduce the excerpts to be rehearsed, while singers were singing with microphones muted. The reproduction in the pilot provided a closer sensation of being singing in a choir than using a piano or another instrumental track.
- ❖ The need to generate **synchronized mixes** from individual recordings of singers. This was possible thanks to the recording possibility of the pilot and the score synchronization functionality. Figure 5.2 shows the list of recordings of one singer in their own rehearsals visualization.

Actions	Piece	Part	Bars	Tempo	Date	Visibility	Audio	My notes	Conductor's notes
📁 ❤️ 👤 🗑️	Oy Comamos Y Bebamos	Soprano	1-21	74 bpm	2021-04-09 10:56:43	👤 Conductor ▾	▶️	📄	-
📁 ❤️ 👤 🗑️	Sus, sus, sus	Soprano	1-59	89 bpm	2021-04-09 10:54:05	👤 Conductor ▾	▶️	✍️	-
📁 ❤️ 👤 🗑️	Oy Comamos Y Bebamos	Soprano	1-21	74 bpm	2021-04-09 08:19:45	🔒 Private ▾	▶️	✍️	-
📁 ❤️ 👤 🗑️	Sus, sus, sus	Soprano	1-59	89 bpm	2021-04-09 08:17:02	🔒 Private ▾	▶️	✍️	-

Figure 5.2. Screenshot of Cantamus app including recordings of a Soprano singer for the pieces Sus, Sus, sus and Oy Comamos y Bebamos. The three top ones were shared with the Conductor for the final mix.

5.1. Pilot Evaluation within the context of Renaissance repertoire

5.1.1. Aim of the evaluation study

For the pilot study we organized a workshop with one of the choirs involved in the last evaluations of TROMPA (La Violeta) to share and review the main functionalities of Cantamus, especially the recording and analysis of the performance functionalities. Apart from the review, this workshop served to gather feedback on the particular repertoire that we recorded with the vocal ensemble Cantoría (*Cantoría repertoire*) and refine the methodology for a larger-scale use. La Violeta worked individually with the two first pieces mentioned before two pieces from the repertoire were recorded with real voice by Cantoría. After the workshop with La Violeta, each singer was given a week to record themselves performing the two pieces on the Choir Singing Pilot, which were later used to generate a virtual mix.

5.1.2. Participants

5.1.2.1 Recruitment strategies

The Violeta choir is the one that has shown the highest level of interest during all its collaboration with the TROMPA Choir Singers Pilot. We believe that the reasons for this high involvement are the interest of the director, the cohesion of the group and the mutual help between the director and members of the board direction. This choir was the first to follow the Choir Singers Pilot training and the first one to have their repertoire on it, so they have had slightly more time than the rest of the choirs. Furthermore, the choir director has used the pilot to complement the part-time vocal rehearsals as an harmonic support. He has also personally helped singers with difficulties to perform the tasks that have been asked to them. Thus, we decided this choir was the best option to evaluate the current status of the pilot and provide insightful feedback. We also asked the UPF (Universitat Pompeu Fabra) choir given the different age rate but this group was inactive due to COVID-19 situation, as many of the students had returned to their countries and the conductor had trouble engaging the singers. Figure 5.3 shows the flier we did for dissemination.



Figure 5.3. Workshop flyer

5.1.2.2 Participant characteristics

The Violeta choir is based in Centelles (Barcelona), central Catalonia. It could be considered a representative choir for the many amateur choirs that exist in Catalonia, due to its musical formation, gender and age: They have a long experience in choral groups, with an amateur level but with an average ability to read music. We collected the data of the members of the choir when we carried out the face-to-face training of the Choir Singers Pilot at the beginning of the project. The intention was to know the demographic characteristics of the group and their musical knowledge. This information was important to assess the use of the Choir Singers Pilot, specifically the performance and analysis functionalities of a specific repertoire, by a typical singer of an amateur choir. Global characteristics can be found in Table 5.1.

Choir	Participants (male/female)	Conductor	Ages	Musical skill distribution (high, medium, low)
Violeta	28 (11 / 17)	Male	35-65	20%, 40%, 40%

Table 5.1 Characteristics of the members of the Violeta choir

5.1.3. Study protocol

Members of Violeta choir already had their own access to the Choir Singers Pilot. For this study, we uploaded the Cantoría repertoire in their profiles. The flow of communication with Violeta was very intense, due to the involvement of the conductor, through WhatsApp, telephone, as well as many calls from the singers for technical issues.

In this session held online, the main features of the pilot were briefly recalled³⁴, with special emphasis on the recording and analysis features. The participants had some tasks to do after the workshop, which consisted of learning, recording their own voice for the 2 given pieces, and exploring the analysis features, with a week of time to do it. This first activity was held with 22 out of 28 members of the choir, and was recorded to make it available to those who could not attend. We already had their direct contacts by email and they have also access to the Choir Singers Pilot.

The recordings of the singers were converted into a virtual choir mix, combining the user recordings with the Cantoría singers, so this mix can be seen as “rehearsal of Violeta with Cantoría”.

The director of Violeta stated that he himself would help some users who did not know how to use the pilot to do the recordings in some face-to-face sessions. After a week of carrying out the first workshop, a second workshop was held to give some premises regarding the interpretation of the pieces, by the director of Violeta himself. It served to expose the problems or setbacks that users had encountered in the use of Cantamus. Unfortunately, this second session could not be held due to the situation of the pandemic, because the director was confined. Most of the explanations were then resolved over the phone and all referred to doubts about the process of recording functionality.

³⁴ https://trompamusic.eu/deliverables/TR-D6.6-Working_Prototype_for_Singers_v2.pdf

5.1.4. Study evaluation outcomes

Once all the tasks had been completed, we shared with the choir the result of the virtual choir with their voices ([link to audio mix](#)). This includes 11 recordings from Violeta from 7 different singers. Some singers sent multiple recordings, and some recordings had to be discarded due to insufficient recording quality (e.g. low audio level, presence of background noise).

This study allowed us to test that Cantamus worked properly, to solve some final technical issues and needs for explanations (e.g. on how to select parts to be rehearsed or recorded), and prepare the pilot for a larger-scale usage as explained in the coming sections.

5.2. Live/on-line participatory concert with Cantoría

5.2.1. Aim of the evaluation study

This action includes a set of activities organized in the context of a concert in Barcelona on April 13, 2021. They included a set of activities including online conferences, a rehearsal and the participation in the live/virtual concert. Registered people had access to Cantamus to study and listen to the scores, record, analyze and share their own interpretation. Due to the regional confinement for the pandemic, the participatory concert was finally organized in an online format, and we later created a virtual choir with the recordings uploaded by the participants. Although this action was first intended just for the TROMPA choirs, given its virtual character we decided to open it to the general Spanish speaking public.

These activities have allowed us to increase the usage of Cantamus on a larger scale, with the Cantoría repertoire of three pieces. Thanks to this large-scale usage we obtain a more concrete and analyzable feedback on how to improve the experience of using the pilot and obtain user recordings and annotations on these specific pieces. The subsequent questionnaire for the participants has reflected their experience on the same scores in terms of usability, unlike the past workshops with individual choirs. This evaluation was intended to obtain data on the most advanced state of Cantamus's functionalities, focused on the process of recording and analyzing individual performances.

5.2.2 Additional functionalities of the Choir singing Pilot

For this evaluation of the Choir Singing Pilot, we implemented two of the identified functionalities in previous evaluations:

- ❖ **Instrumental track:** it is common that choir repertoire contains an instrumental accompaniment by orchestra or piano. This has been extensively requested in previous user studies. We implemented a new audio track to the CSP, allowing choirs to upload an instrumental track, in the form of an audio MP3/WAV file. This instrumental track is perfectly synchronised in time with the MusicXML score.
- ❖ **Practice tracks** (expressive performances with varying tempi): in addition to the functionality of synthesising scores with artificial voices, some choirs requested the possibility to upload already available practice tracks³⁵. We implemented this new functionality with Cantoría

³⁵ Practice tracks are audio recordings for each part (e.g. four tracks in the case of SATB scores) that choir conductors share to choir members. These tracks can typically be a singing recording or also a piano reference recording with the melody.

tracks, in order to upload one audio file per part. We carried out manual score alignment with a time-varying tempo, as we find in real expressive performances.

5.2.3. Participants

5.2.3.1 Recruitment strategies

When we decided that the participatory concert in collaboration with Cantoría would be open to the general public, we considered various communication actions to disseminate it. The participatory concert was also held by students of Escola Superior de Música de Catalunya (ESMUC). We contacted ESMUC and they gave us facilities for the online workshops and contributed to dissemination through their press office, as well for the dissemination of the flyer (see Figure 5.4). Cantoría is very well-positioned in social media, especially in Twitter and Instagram, so we decided to use the @TrompaMusic Twitter account to disseminate the actions.

The media echoed the initiative and we had the following inputs:

- ❖ Pompeu Fabra University [news](#).
- ❖ ESMUC [news](#):
- ❖ Emilia Gómez presented TROMPA and the Choir Singing participatory concert at Spanish [National Radio](#), Radio Clásica, Longitud de Onda
- ❖ Emilia Gómez presented TROMPA at a program on AI and music at the [Spanish Radio](#);
- ❖ Emilia Gómez presented Trompa at Popap program on [Catalunya Ràdio](#) (43');
- ❖ Jorge Losana, conductor of Cantoría, presented the participatory concert at Assaig general program on [Catalunya Ràdio](#);
- ❖ "Cantāmus" herramienta en línea que permite aprender el repertorio de los cantantes", Plaza Pública, [Regional Radio from Murcia](#), Spain;
- ❖ Social media dissemination³⁶. The maximum diffusion on the Twitter channel was a tweet during the day of the concert, as shown in Figure 5.5
 - 3.389 impressions;
 - 100 engagements.

³⁶ <https://twitter.com/TrompaMusic>



Figure 5.4: Participative concert flyer

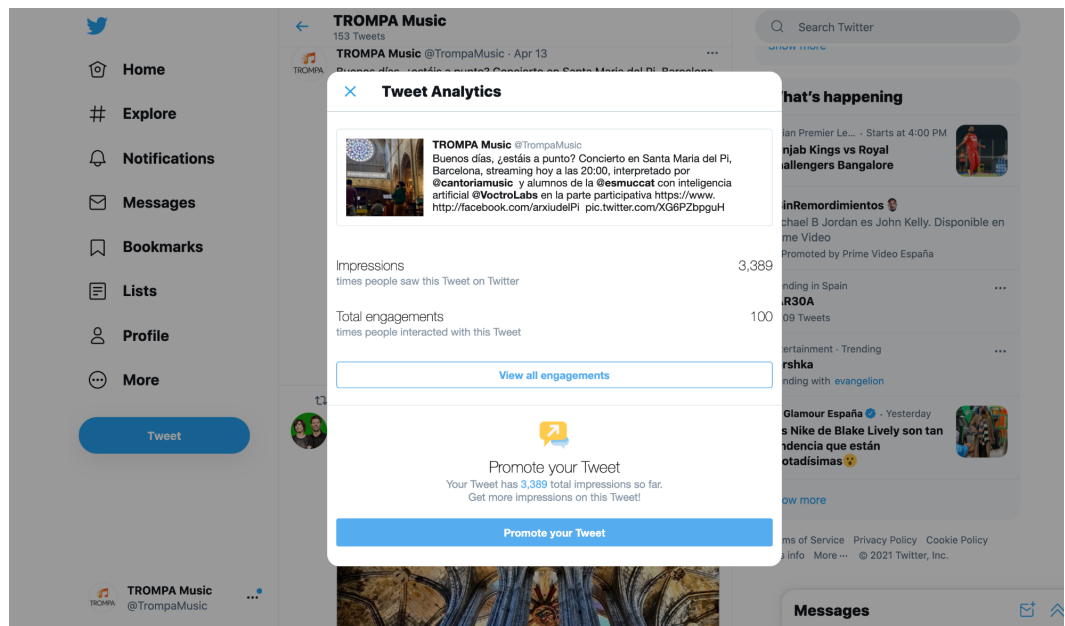


Figure 5.5: Maximum interactions on the Twitter channel @trompamusic

The data used to disseminate the activity included information from different choir entities related to TROMPA:

- ❖ TROMPA choirs;
- ❖ Cantoría database (of the previous participatory online concert they did with the “More Hispano” project);
- ❖ FCEC database from the Catalan Federation of Choirs;
- ❖ TROMPA members contacts.

5.2.3.2 Participant characteristics

Information about this activity was sent to all TROMPA choirs, Cantoría contacts (who had already carried out a participatory format at Christmas on the piece *Sus, sus, sus*), and the general Spanish-speaking public, as all the information was generated in Spanish and Catalan.

We got a total of **130 registrations** to the activity. We can say that the main common characteristics of the participants are their interest in music and their active participation in this type of participatory actions. With the TROMPA choirs we were able to address amateur Catalan choirs, aged between 25 and 68 years. With Cantoría's contacts, we expanded the community to the Spanish and Latin American choral world. A group of singers from Latin America showed a lot of interest in the Spanish Renaissance music and Cantoría's actions. They also have been very active in our actions.

The subsequent questionnaire showed that most of our participants, 84%, used a computer or a laptop, which was recommended over mobile devices, due to the size of the screen, as shown in Figure 5.6

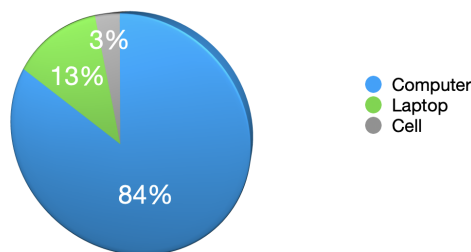


Figure 5.6: Device used to test the application

In Figure 5.7, we observe that most participants are over 46 years old, and around 80% have sung for more than 5 years in a choir. This is the expected population for our users, which include senior people and also amateur singers.

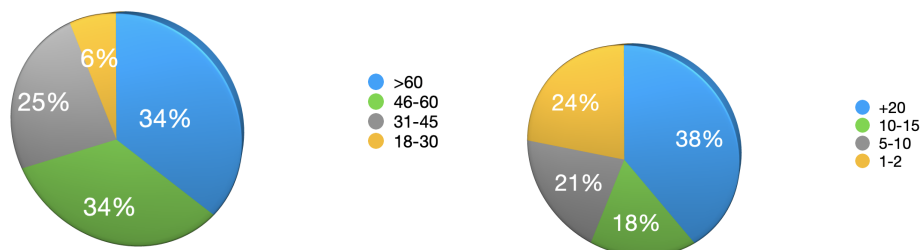


Figure 5.7: Age of the participants (left) and years of participation in a choir (right)

In Figure 5.8, we observe a majority of women (72% of soprano and alto), which is commonly found in choirs.

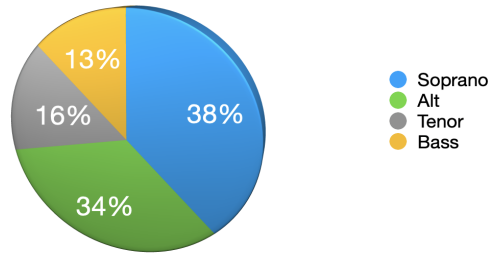


Figure 5.8: Type of voice

Finally, we observe in Figure 5.9. that many users declare they have advanced knowledge on music and most of them can read a score. This will be coherent with the results of the visualizations as we will comment in the coming sections.

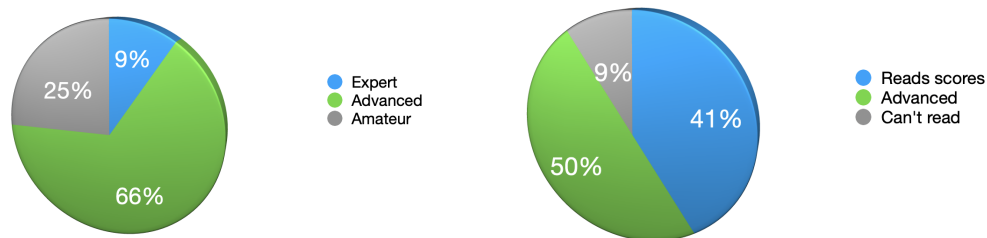


Figure 5.9. Musical knowledge level (left)³⁷ and music reading level (right)³⁸.

Figure 5.10 illustrates the active sessions on the Cantamus platform for each day of the study. We observe that from the 130 registered participants, only 99 of them actively used the Cantamus app in the end, and the maximum number of joint connections was 42 (people accessing the same time in the same day).

³⁷ **Expert:** Solid musical training. **Advanced:** some years of formal music education, **Amateur:** No formal music education.

³⁸ **Question:** Can you read scores? **No:** I can't **Reads scores:** I can interpret them, but not fast enough to sing as I read **Advanced:** Yes, I can read sheet music fluently and sing as I read



Figure 5.10. Active sessions in the Cantamus platform during the week of the activity.

5.2.4. Study protocol

Once registered, we send a second email with the required information for the study, including the access information to the Choir Singers Pilot Cantamus, musical scores, links for accessing the complementary activities and link to the concert in streaming.

All the new users of Cantamus were manually created, importing a list of names and emails to the platform. The users were also provided with an email address to attend to possible questions, proposals, unforeseen events, etc.

The set of participatory activities were attended by the participants in the following order. We also incorporate a summary of attendance to each of them:

1. Workshop about the Choir Singers Pilot, Cantamus, was attended online by 60 people (and later viewed by 360 people on the Escola Superior de Música de Catalunya (ESMUC) Youtube channel³⁹). In this workshop, TROMPA partner VL provided a description of the tool and answered questions about its usage. Figure 5.11 shows some photos of the workshops.
2. Musicological conference by Professor Maricarmen Gómez Muntané, emeritus professor at the Universitat Autònoma de Barcelona and specialist in the target repertoire. It was watched online by 50 people (with 275 total views on the Escola Superior de Música de Catalunya (ESMUC) Youtube channel⁴⁰). In this conference, the participants were introduced to the musicological characteristics of the pieces of the repertoire and their historical context.
3. Rehearsal with the members of Cantoría and TROMPA: this rehearsal included some choir singing rehearsal, brief discussion about the tool and the TROMPA project. It was held with 54

³⁹ <https://t.co/7V1kRPEZi0?amp=1>

⁴⁰ <https://www.youtube.com/watch?v=IVl-i1OFs8Q>

people. During the rehearsal, the Cantamus tool was used to reproduce the pieces, since online conference platforms are not prepared to carry out a rehearsal with the participants singing. The video recording is available on YouTube⁴¹. Figure 5.12 shows a moment of the online rehearsal.

4. A participatory concert which was finally not possible due to COVID, but the Basilica de Santa Maria del Pi offered it by streaming. Unfortunately, various technical problems caused the connection to be lost in the middle of the concert and the participants could not follow all of it.
5. Elaboration of a final virtual choir, made with the rehearsals recorded by participants in the Choir Singing Pilot. We obtained, for one of the pieces, 37 singers: 33 workshop participants (16 soprano, 10 alto, 2 tenor, 5 bass) and 4 Cantoría members. For the other piece it included 34 singers: 30 workshop participants (13 soprano, 11 alto, 2 tenor, 4 bass) and 4 Cantoría members. The final mixes can be found online⁴². A screenshot of the score *Oy comamos y bebamos* is presented in Figure 5.13.
6. During all of the process, we gathered comments and feedback through the YouTube chats and questions on the online activities. After all the activities, we sent an evaluation form to the participants to gather their feedback in a more formal way, in Catalan and Spanish.

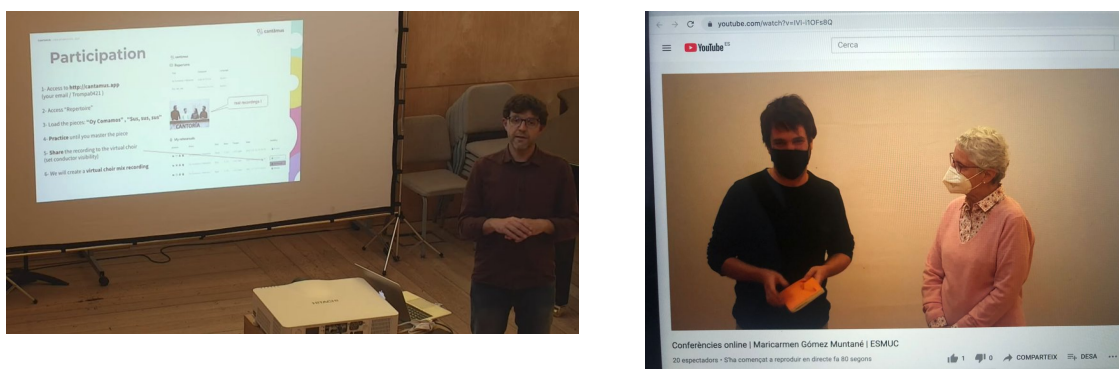


Figure 5.11. Presentation of TROMPA and the Choir Singing Pilot Cantamus (left) and musicological conference (right), in the context of the participatory concert

⁴¹ <https://www.youtube.com/watch?v=yWrP96Y1BYw>

⁴² <https://voctrolabs.bitbucket.io/virtualchoir/>



Figure 5.12. Virtual rehearsal of the participatory concert (blurred for personal data protection)

The screenshot displays the Cantamus app interface for a virtual rehearsal. On the left, a 'Voices' panel allows users to select and adjust volume for Soprano, Alto, Tenor, and Bajo, as well as an Organ part. The main area shows the musical score for the piece 'Oy Comamos Y Bebamos'. The score is written for Soprano, Alto, Tenor, and Bajo voices, and Organ. The lyrics are: 'Oy co - ma - mos y be - ba - mos, y can - te - mos y hol - gue - mos, que ma - ña - na a - yu - na - re - mos. Por on - ra de San An - true - jo, pa - ré - mo - nos oy bien'. The interface includes playback controls like 'Listen' and 'Practice', and a 'Bar selection' tool.

Figure 5.13. Screenshot of the Cantamus app with one of the pieces of the repertoire

5.2.5. Study evaluation outcomes

In this subsection we present a summary of the evaluation results. In our evaluation, users were asked to rate various elements of the CSP from 1 to 5. In most cases (score display, piano roll display, and overall impression, difficulty in the recording process, and voice analysis) these items are rated 4 or 5. The use of the recording analysis tools shows some difficulty, and there is a quite high degree of users who have not used any of the two analysis tools (piano roll or voice analysis rating). We hypothesise that it is due to the high musical knowledge of the audience.

In Figure 5.14 we observe how users prefer score display over the piano roll view. Our hypothesis is that it is the format usually used in choirs by people with high musical knowledge, so it is more familiar for them. In addition, people are not familiar with the use and visualization of singing ratings, illustrated also in other questions that will be mentioned below.

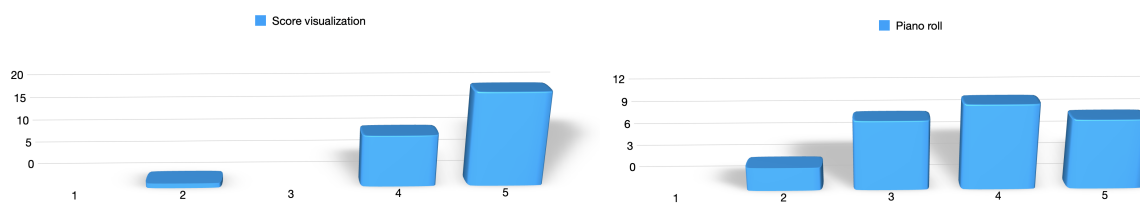


Figure 5.14. Rating of the score display (left) vs piano roll view (right)

Figure 5.15 shows that the pieces under study were somehow challenging for the users, who then used the tool to support this learning process. Due to the nature of the pieces, we observe that the intermediate voices (alto and tenor) are more difficult in these compositions compared to the soprano and bass.

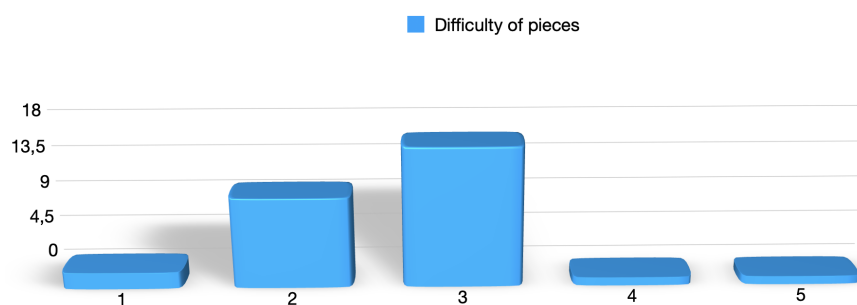


Figure 5.15. difficulty of the 2 pieces performed

Figure 5.16 shows that users found it slightly more difficult to understand the analysis rating part than the recording process. We think it is due to the fact that most users are already familiar with voice recording functionalities, but not much with singing performance rating algorithms, so it is difficult for them to use it for their own rehearsal.

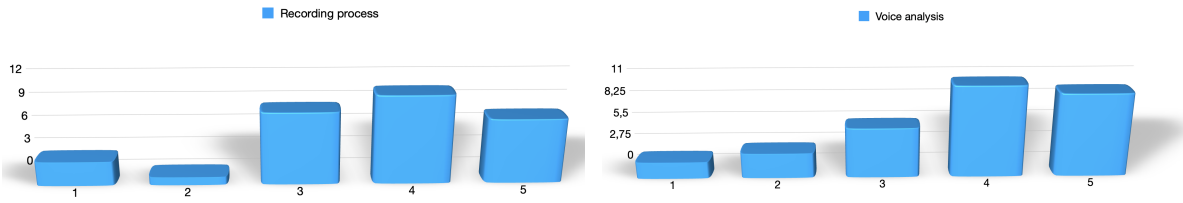


Figure 5.16. difficulty of recording process (left) and the voice analysis rating (right)

Figure 5.17 shows that many users didn't use the performance rating in the rehearsal, and some looked at the coloured notes or the coloured notes and the piano roll.

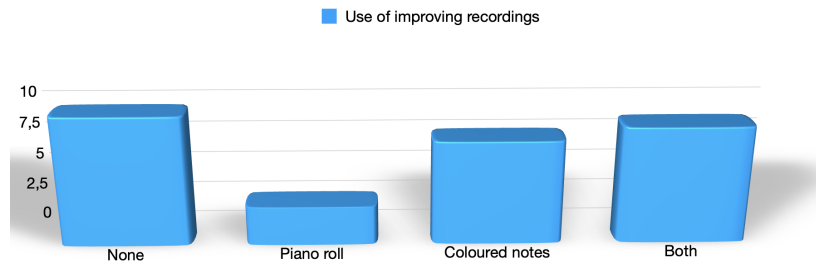


Figure 5.17. Use of improving recording features

Finally, we observed that users were generally satisfied with the pilot, as we see in Figure 5.18.

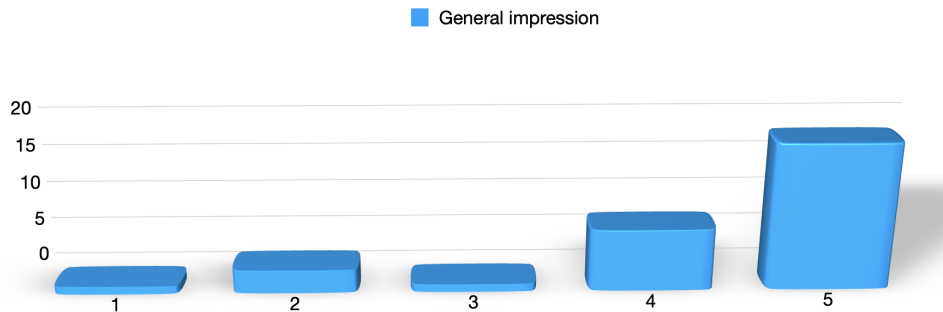


Figure 5.18. General impression on the Choir Singing Pilot

In addition to this evaluation form, we solved some problems by email, summarized by subject in Table 5.2:

ISSUE	FREQUENCY
Usage of the recording functionality	15
Usage on the analysis/rating functionality (piano roll)	5
How to share our rehearsal/recording with the conductor	3

Access problems	5
Problems with computer/navigator	3
Other ⁴³	4

Table 5.2. Table of most common problems solved by email

The form also included some questions about future exploitation of the Cantamus app. On one hand, we asked about the repertoire the users would like to be available, as we can see in Figure 5.19.. On the other hand, Figure 5.20 shows the preferences of users on behalf of the type of subscription. In this last question, 52% answered they would like a subscription as a member of a choir with the choir's repertoire available in the app.

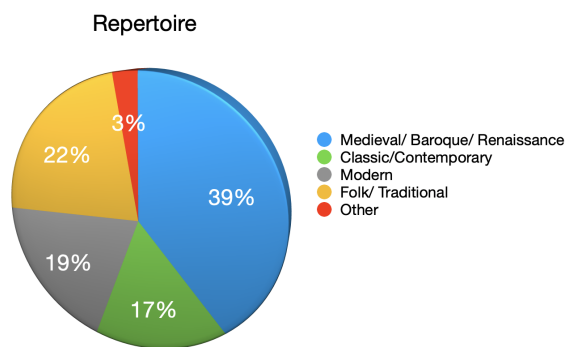


Figure 5.19. Repertoire preferences for the future

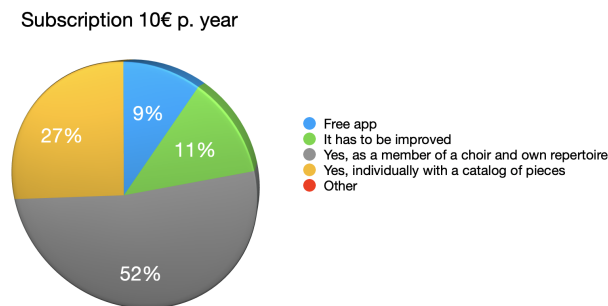


Figure 5.20. Future subscription preferences

5.3. Conclusions

Two important conclusions have been drawn from the evaluation of the Choir Singers Pilot (CSP) prototype by users: i) there is a need and interest among the choir singers for using technologies like the CSP that allow keeping the activity in the current COVID-19 situation; ii) the current version of the CSP is appropriate in terms of provided functionalities, and requires some usability

⁴³ The rest of the questions were varied and heterogeneous, such as some people who asked for the MIDI file to listen to, others who made proposals for improvements and some about a doubtful note in the reference recording.

improvements to be viable as a commercial end-user solution. It is worthy to mention, as a general conclusion from all evaluation exercises for the pilot in the project, that we observed that the algorithms incorporated in the pilot, e.g. singing voice rating and synthesis, were not intuitively used by the singers in the first sessions, due to the fact that these are novel with respect to the other tools they use. As a consequence, we need to closely work with conductors and singers to incorporate these algorithms more into the rehearsal, as it was the case, for instance, for the Violeta choir, which made extensive use of the platform.

In parallel to the Cantoría activities, we have been contacting and being contacted by other choirs in Spain and also internationally (e.g. Margaret's Choir, Winnipeg, Canada) that have shown interest in using the prototype in a Pilot test setup. In this case, we are taking advantage of the singing synthesis capabilities to render custom scores. These new studies will extend over the end of the TROMPA project.

As we explain in the **Deliverable 7.3-2 - Exploitation Plan** deliverable⁴⁴, the above conclusions have encouraged VoctroLabs, as one of the consortium's SMEs, to keep working on the prototype in view of a commercialization during the second half of 2021.

6. Music enthusiasts

Under the Music enthusiasts use case, several user evaluations have been conducted during the last year, and they were reported in **Deliverable 6.7-2 Working prototype for music enthusiasts**⁴⁵ and **Deliverable 6.8 - Mid-term evaluation**⁴⁶. Section 6.1 describes the second iteration of the user evaluation described in D6.7-2, following the format of an online one-week contest to evaluate the pilot in a real setting. Additionally we conducted a long term campaign to assess the refinements made in the previous campaigns in terms of the collected annotations and the inter-rate agreement. The results of the long-term campaign are described in Section 6.2. Additionally, in Section 6.3 we present an analysis of the platform usability based on the user behavior metrics implemented within the platform since april 2020 (first workshop with the working pilot). It is worth mentioning that the platform was released to production on april 2020, and all the new features and refinements have been released in production. Hence, the number of registered users and collected annotations is higher than the total of participants of the workshop and the user evaluation studies conducted.

6.1 Second contest: Music from West Africa

6.1.1. Aim of the evaluation study

The aim of this evaluation study was to evaluate the usability and workflow of the pilot in a real setting (participants using the ME platform by their own with their own devices), as well as to determine the impact of the implemented incentives (e.g. scoring system, contributors' ranking, music recommender system based on emotional content) and the quality of the annotations. Likewise, the evaluation study allowed to assess the scope of the dissemination mechanisms

⁴⁴ This deliverable is confidential to the consortium only.

⁴⁵ https://trompamusic.eu/deliverables/TR-D6.7-Working_Prototype_for_Music_Enthusiasts_v2.pdf

⁴⁶ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

available, e.g. mailing lists and social networks. This evaluation study focused on user behavior data collected through the platform and the data collected from questionnaires. This evaluation study is complementary to the one presented in the **Deliverable 6.8 - Mid-term evaluation**.

6.1.2. Participants

6.1.2.1 Recruitment strategies

Recruitment strategies have been implemented following the ones of the previous contest. UPF networks (Twitter, mailing lists, etc.) were the main recruitment strategy. TROMPA social networks were also used to disseminate the contest (Figure 6.1). In addition, participants who participated in the previous contests, already registered on the platform and opt-in in receiving updates about the platform, have been contacted through mail. At the time of the contest, the pilot ran in English and Spanish, so the call for participation messages were disseminated in both languages.

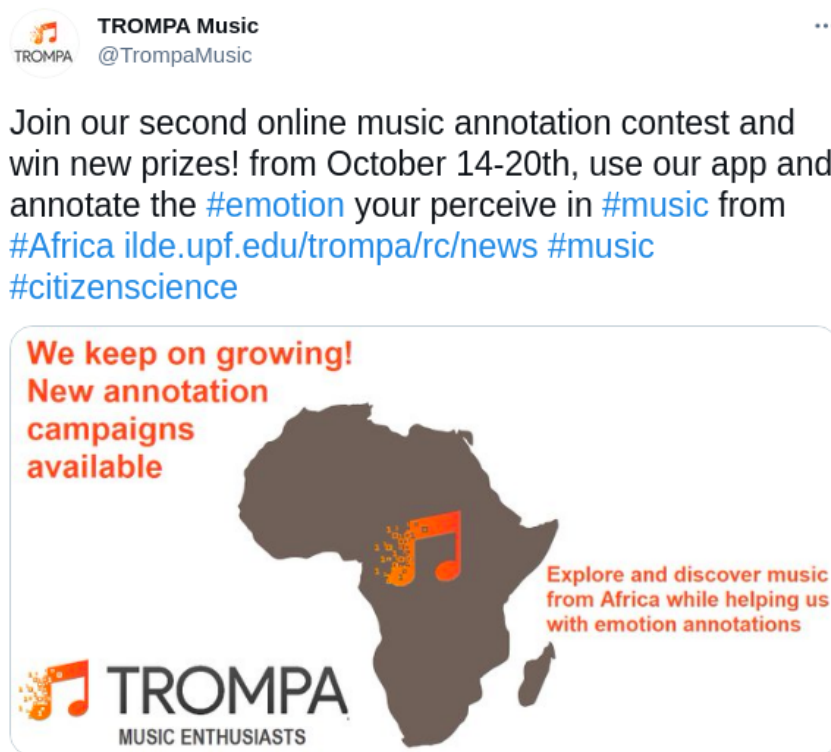


Figure 6.1. Example of a dissemination tweet promoting the contest.

6.1.2.2 Participant characteristics

- ❖ Participants were English and/or Spanish and/or Catalan speakers. However, only four participants added personal information.
- ❖ During the contest period, 23 participants generated 655 annotations from 202 songs.

6.1.3. Study protocol

The protocol implemented in this second campaign follows similar guidelines as defined for the first contest (see **Deliverable 6.8 - Mid-term evaluation**⁴⁷, Section 6.3). Participants had to annotate as many songs as possible in order to obtain an external reward (Bandcamp gift cards). The winners of the contest were defined using the same scoring system of the previous contest. General rules for the contest were defined as follows:

- ❖ Participants must login in the pilot. In order to register, users must accept the TERMS OF USE of the platform. The information sheet is presented to the user, where detailed information about the collected data, use of this data for research purposes, as well as the privacy policy of the pilot is described.
- ❖ Once they were registered, they were able to annotate.
- ❖ Participants must complete the Tutorial campaign in order to access the rest of the campaigns.
- ❖ Participants must complete at least one of the available campaigns (different from the Tutorial campaign).
- ❖ In case of a tie, the winner of the prize has been determined as follows:
 - Highest amount of valid annotations done during the contest period.
 - Highest amount of completed campaigns during the contest period.
 - Highest amount of annotations during the same access to the platform.
 - If the tie persists, the prize will be awarded by lottery.

In addition to the prizes given to the two participants who obtained the highest score in the ranking (First place: 50 euros Bandcamp gift card, Second place: 20 euros Bandcamp gift card), a third prize (20 Euros Bandcamp gift card) has been given by lottery among the participants who completed at least one campaign and additionally filled the availability survey⁴⁸.

Previous to the contest, there were 84 songs that were annotated by participants of previous contests described in **Deliverable 6.1 - Final Mock-ups Testing**⁴⁹ and **Deliverable 6.7-2 - Working prototype for music enthusiasts**⁵⁰, contained in two campaigns (Campaign 1 and Campaign 2), and 4 campaigns focused on different types of music in Spanish and Portuguese languages (see deliverable D6.8). For the contest, 6 additional campaigns were incorporated within the platform so the participants could explore different types of music from West Africa.

In the first contest we selected Spotify gift cards as external reward, since it is the most used streaming platform among the target audience of the pilot. Nevertheless, for the second contest we selected Bandcamp gift cards since ME Pilot goals is to help participants to discover new music, and Bancamp is a promotional platform for independent artists and new music.

6.1.4. Study evaluation outcomes

Based on the collected data during the second contest, the following conclusions were drawn:

⁴⁷ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

⁴⁸

https://docs.google.com/forms/d/e/1FAIpQLScaAkdN0Gv4W71MtMoPAXIwl_0m6UqOofG8e8MJeeTK2LF1fg/vi ewform

⁴⁹ This deliverable is confidential to the consortium only.

⁵⁰ https://trompamusic.eu/deliverables/TR-D6.7-Working_Prototype_for_Music_Enthusiasts_v2.pdf

- ❖ Figure 6.2 shows the distribution of annotations across the different songs to be annotated. As stated in **Deliverable 6.8 - Mid-term evaluation**⁵¹, we found that new users would initially make annotations from the initial campaigns and would abandon the task before listening to new music from West Africa.

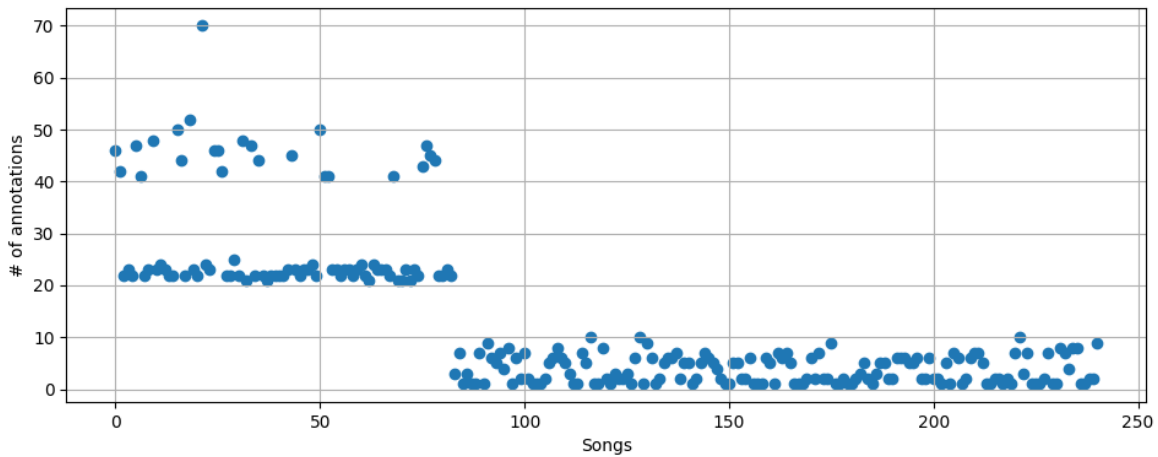


Figure 6.2. Annotation distribution after the second contest.

- ❖ We evaluated the reliability of the collected data using Krippendorff's coefficient α to understand the importance of inter-rater agreement on the collected annotations. In summary, we obtained: (1) $\alpha_{Arousal} = 0.505$, (2) $\alpha_{Valence} = 0.364$, and (3) $\alpha_{Emotion} = 0.192$. Additionally, we find a marginal increment of agreement regarding arousal, probably due to the implementation of the tutorial which explains the reasons for annotations.
- ❖ Due to the lower rate of participants who completed the personal information, we modified the platform workflow to redirect participants to the user settings when they register in the platform or if they haven't completed yet all the information.
- ❖ The second contest had similar results than the first contest. The number of gathered annotations and the registered participants was similar. Nevertheless, during the second campaign a couple of "great" contributors emerged: four users generated over 100 annotations during the contest. This is a normal behavior in online communities, where a reduced number of contributors generate the greatest number of contributions. To tackle this, we aggregated a "weekly score" to avoid demotivating participants with lower scores.
- ❖ Regarding the results in terms of gathered annotations and registered users, both campaigns had similar results. Thus, both external rewards motivate contributions in a similar manner.

6.2 Long-term campaign: Music from Latin America

6.2.1. Aim of the evaluation study

In this campaign, we formulated a different approach: a daily playlist with 20 songs over a period of 27 days. Since several participants rated the initial songs from previous campaigns, we allowed only the songs from this campaign to be visible. We added songs from the following countries: Argentina,

⁵¹ https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, Mexico, Panama, Peru, Puerto Rico, Uruguay, and Venezuela. In this approach, a longer time to collect annotations could result in more participation. Since we expect the response diversity to be high, we attempt to collect an enriched and varied set of responses. To this extent, our platform has been translated into English, Spanish, Italian and Dutch. We also provided a link to an usability survey that participants could complete voluntarily.

6.2.2. Participants

6.2.2.1 Recruitment strategies

We extend previous strategies for recruiting participants, based mainly on UPF-MTG, UPF-TIDE and TROMPA networks, promoting the long-term campaign on the following platforms (Figure 6.3):

- ❖ Twitter⁵² (English / Spanish):
- ❖ Instagram⁵³ (English / Spanish):
- ❖ Reddit⁵⁴ (English)
- ❖ TROMPA website⁵⁵ (English):
- ❖ Banner in Muziekweb website (Dutch / English / French)
- ❖ UPF website⁵⁶ (English / Catalan): link
- ❖ La Vanguardia website⁵⁷ - (Spanish daily newspaper) (ES): link

Using this strategy, we have been able to reach a wider audience, enabling us to collect a varied set of responses.

⁵² <https://twitter.com/TrompaMusic/status/1356886967057870849>

⁵³ https://www.instagram.com/p/CK39E2_gDU-

⁵⁴

https://www.reddit.com/r/CitizenScience/comments/ld2rlb/survey_looking_for_music_enthusiasts_to_study/

⁵⁵ <https://trompamusic.eu/node/131>

⁵⁶

https://www.upf.edu/web/focus/noticies/-/asset_publisher/qOocsyZZDGHl/content/id/242646593/maximized#.YIGoYYOA5ph

⁵⁷

<https://www.lavanguardia.com/vida/20210203/6220676/upf-inicia-campana-sobre-sensaciones-provoca-musica.html>



Figure 6.3. Banner used for dissemination of the long-term campaign.

6.2.2.2 Participant characteristics

- ❖ Participants were English, Catalan, Spanish, Italian and Dutch speakers. For the first time we reached participants from other countries such as Portugal, Argentina, Colombia, and the United States.
- ❖ During the contest period, 26 participants registered in the platform and 183 annotations from 70 songs were generated by 23 participants.

6.2.3. Study protocol

We change the protocol of this campaign, given its different design and methodology in presenting the songs to be annotated. In fact, no rewards have been given to participants for participating in this campaign. Additionally, instead of presenting all the list of songs at the beginning of the campaign, we add periodically new lists (i.e. a new one every day), to engage participants in discovering new content every time they access the platform. However, we preserve some parts of the protocol implemented in the previous contest:

- ❖ Participants must login in the pilot. In order to register, users must accept the TERMS OF USE of the platform. The information sheet is presented to the user, where detailed information about the collected data, use of this data for research purposes, as well as the privacy policy of the pilot is described.
- ❖ Once they were registered, they were able to annotate.
- ❖ Participants must complete the Tutorial campaign in order to access the rest of the campaigns.

With regards to the content, in this campaign we choose to remove the songs included in the previous campaign, to give the possibility to the participants to focus exclusively on the new lists provided. These lists contain 540 songs from Latin America, divided by country of origin, namely Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, Mexico, Panama, Peru, Puerto Rico, Uruguay, and Venezuela.

6.2.4. Study evaluation outcomes

Based on the collected data during the third contest, the following conclusions were drawn:

- ❖ Figure 6.4 shows the distribution of annotations across the different songs to be annotated. Since in the previous campaign, the users abandoned the task before arriving at new music, we presented only the new music to our participants. Hence, several annotations were collected for the tutorial (first 4 songs), and less for the rest.

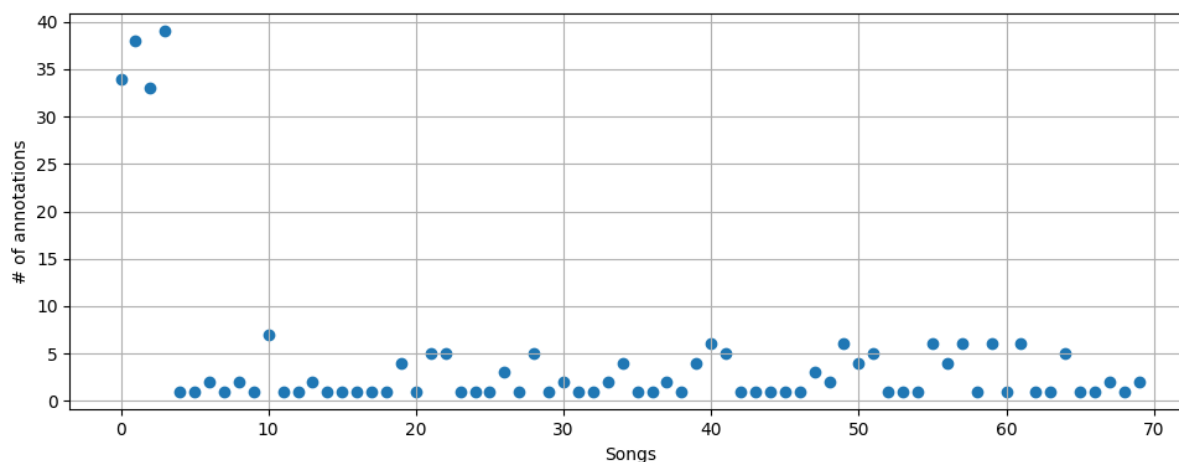


Figure 6.4. Annotation distribution after the third contest.

- ❖ Since we introduced the tutorial for the previous campaign, participants can now select the reasons of annotation from a curated list of reasons. Thus, it is easier to see the effect of psychophysiological associations with emotion words, as seen in Figure 6.5.

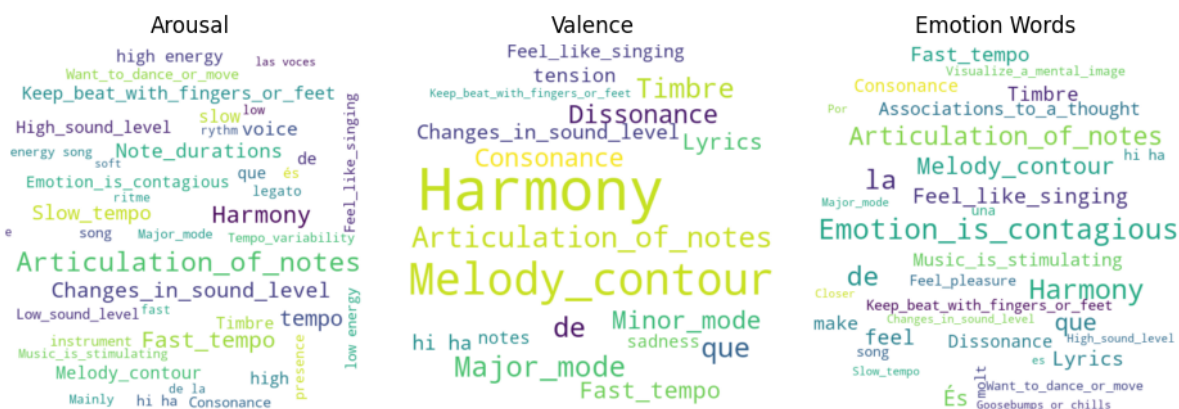


Figure 6.5. Word clouds with reasons for annotation.

- ❖ For each recommendation, we display details regarding tempo, tonality (major or minor), and danceability extracted with the Spotify API [14] (Figure 6.6). For example, a song with 120 BPM, major mode, and 70% danceability will be possibly assigned to the first quadrant of AV space (positive arousal and valence) - reinforcing explanations regarding musical properties of emotion.

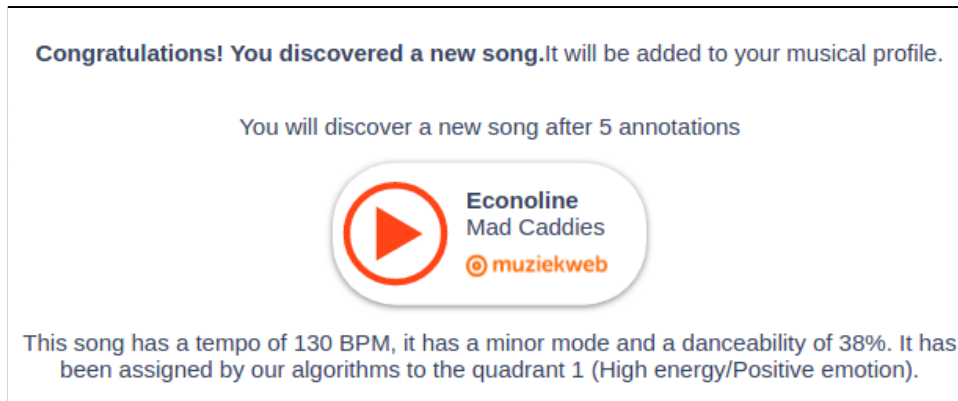


Figure 6.6. Displayed recommendation received after five annotations, including the explanation of some musical features.

23 participants answered the usability questionnaire. Regarding the usability, the platform obtained an average SUS score of 72.9/100, which is the same average value as the one obtained during the previous studies. In fact, results of the t-test showed that there are no significant differences between the results of both versions ($t = 0:003$, $p\text{-value} > 0:05$). Additionally, participants were asked about their perception regarding the music they discovered through the platform (Figure 6.7), and results suggest that participants discovered completely new music when annotating (NM1) since most of it was music that they do not tend to listen to (REC1). Still, the recommendations require refinement to make them more appealing for participants (NM2, REC2). These results suggest that it is still necessary to improve and enhance the incentive mechanisms.

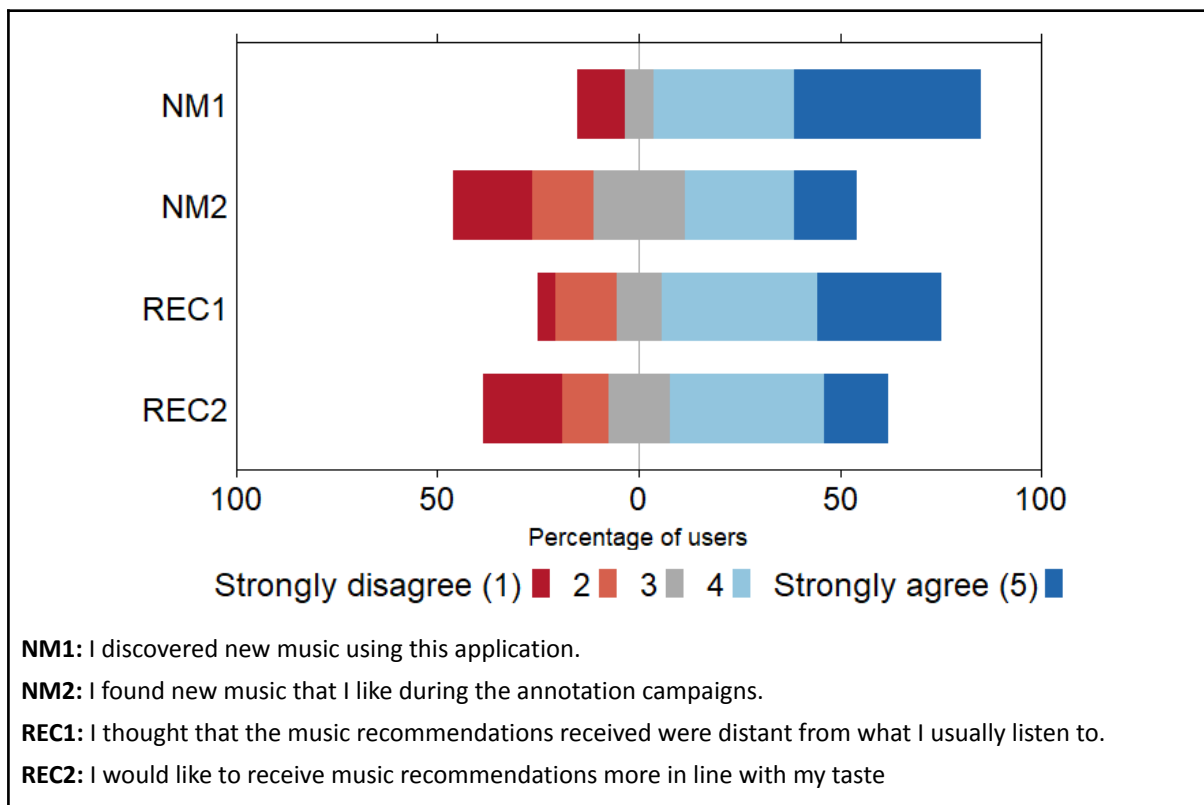


Figure 6.7. Perception of users about the musical recommendation and the discovered music.

6.3 User behavior analysis

Since the first released version of the platform (April 2020), log data have been collected. After the second online contest described in Section 6.2, we gathered additional metrics such as the click counts for each section of the platform. Figure 6.8 presents the results obtained from the collected data. In Figures 6.8.1 and 6.8.2, the peaks represent each of the four user evaluations performed (online workshops in April and May 2020, first online contest in July 2020, second online contest in October 2020 and the long-term campaign on February-March 2021). Based on the results of the analysis, the following conclusions were drawn:

- ❖ As appreciated in Figure 6.8.1, the number of gathered annotations in the long-term campaign was less than the gathered annotations during other campaigns. It is noteworthy to mention that there was no external reward during this long-term campaign, and a lower contribution rate was expected. Opposite to this result, the number of registered users increased (Figure 6.8.2), but not all of them were engaged enough to contribute. This suggests that the recruitment strategies were effective, but the pilot still requires several refinements to engage long term participation.
- ❖ Figure 6.8.3 reveals an expected behavior within any online community. More than 80 participants have less than 10 annotations, while only 20 participants have more than 80 annotations (Pareto principle). Regarding the annotations per song, most of the songs have less than 10 annotations (Figure 6.8.4).
- ❖ Figure 6.8.5 shows the density of the time spent during an annotation. This result suggests that most of the participants listen to the whole music excerpt to perform the annotation (the highest density is around 35.6 seconds). This means that participants understood the scoring system and annotations tend to be more accurate since they are not based only in the first few seconds of the excerpt.
- ❖ The use of the platform has shown that participants are not making use of the 'About us' section and the help menu. This means that these sections should be refined to increase participants' understanding of the project goals, as well as to improve the training phase.

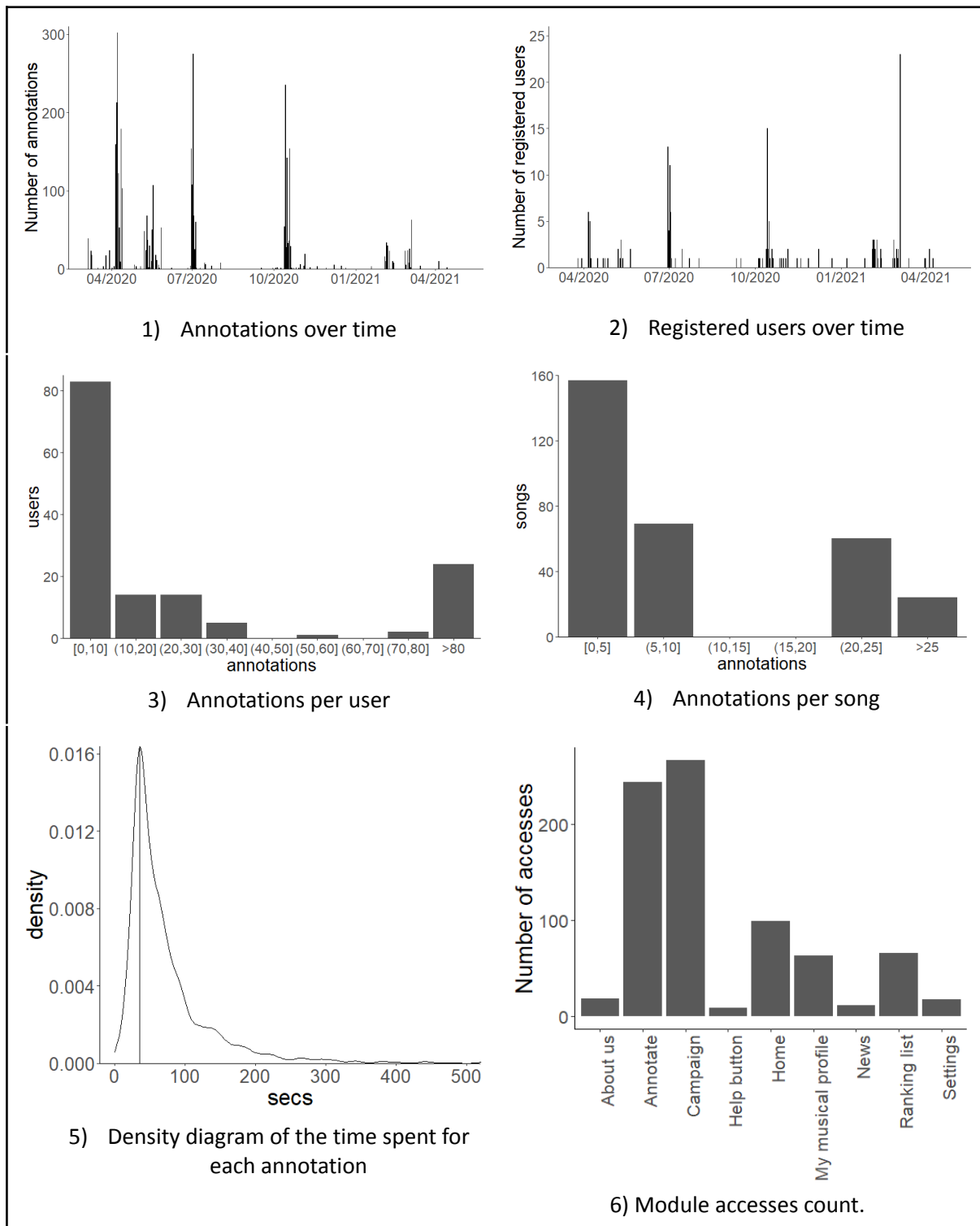


Figure 6.8. Results of user behavior data collected through the ME pilot.

7. Conclusion

In this deliverable, we reported on the evaluation studies, ran on the prototypes that came out of the five use cases. For each of the use cases, multiple of such studies were run, demonstrating iterative progress and increasing versatility of the prototype outcome.

With the unexpected COVID-19 crisis, engaging audiences and running user studies had been more challenging than foreseen at the start of TROMPA. As a consequence, many of the presented studies have been conducted in smaller-scale settings than the project had originally intended. While the crisis-induced online and remote working conditions may technically have made the studies more globally accessible (which also is demonstrated by international participants having joined in several of the user cases), at the same time, we could notice the crisis also led to screen fatigue and lower motivation with our user audiences, which may explain while even large-scale recruitment did not always manage getting large audiences back.

Nonetheless, for each of the use cases, we managed getting in touch with relevant and representative user audiences, who gave very valuable feedback to our work. Several times throughout the TROMPA project (with RCO orchestra members during mock-up studies for the orchestra use cases, with mdw students during the instrument player use case, and with choir singers considering algorithmically prepared functionalities), digital tooling was seen as imposing an extra learning curve, that participants would not always be willing to invest in, considering their daily practice. This is a realistic extra obstacle towards buy-in at larger scale: digital innovation has not necessarily been embraced yet in the classical music communities, even though the COVID-19 crisis did push in favor of this.

Both in case of the Orchestra use case and the Choir singers use case, the enthusiasm of an ensemble's leader for the prototype had great influence on an ensemble's overall engagement and involvement (which we noticed with the AJSO orchestra and the La Violeta choir). Having such engaged champions in leadership roles will be very important for getting digital innovation further out and adopted in the field.

At the same time, many participants in our use case evaluation studies reacted enthusiastically to provided functionality, and did see future promise in our prototypes. Therefore, beyond the lifetime of the TROMPA project, it will be worthwhile to further develop and improve the prototypes. As soon as circumstances will have normalized after the crisis, it will be interesting to revisit the user studies under more ecological conditions.