ONLINE SURVEY ON USABILITY AND USER EXPERIENCE OF MUSIC NOTATION EDITORS

Matthias Nowakowski
Center of Music and Film Informatics
Detmold University of Music
Detmold, Germany
matthias.nowakowski@hfm-detmold.de

Aristotelis Hadjakos
Center of Music and Film Informatics
Detmold University of Music
Detmold, Germany
aristotelis.hadjakos@hfm-detmold.de

ABSTRACT

Although one can see a certain convergence between the interaction designs of different notation editors, there is no general consensus or standard since new interaction paradigms keep appearing with most major software updates or new products. In this paper, we present the results from an online survey ($n = 138$) with standardized usability and user experience questionnaires. The users of digital notation editors were asked to fill out the *System Usability Score*, the *AttrakDiff2* and the *Liveness* questionnaire. This provides insights into domain specific design problems with the goal to inform the design of future interfaces. Almost all music notation editors show clear deficiencies in overall usability. Furthermore, a detailed examination of the obtained metrics show specific dependencies of individual qualities, which are helpful to conduct further qualitative research.

1. INTRODUCTION

Usability is one of the key topics of human computer interaction (HCI) research. It describes the property of a system to help a user achieve goals effectively, efficiently and satisfactorily [1]. In the early days of HCI research much fundamental work was done on the psychological and motorical basics of using text editors and graphical user interfaces (GUI) and thereby also determining their efficient use [2, 3, 4]. In music, the digital positioning of graphical objects such as notes, staves, etc. was already problemized since the 1960s [5, 6, 7]. This was not framed as a topic of usability, but rather of automatizing score editing, as demanded by formatting or engraving and contemporary music notation practices [8]. There was no need for music specific interaction paradigms since the interaction was based on text editing.

Before the first graphical user interfaces (GUI) for music notation were developed, GUIs were used to create electronic music and were seen as a creative tool. Novel and more abstract musical representations could be employed, suitable for interactions via mouse and keyboard. First usability considerations in this regard where made in the mid-1980s by analyzing digital workflows [9], describing interactive graphical environments for computer assisted composition [10] and so essentially proposing ideas for a visual programming languages based on parametric manipulations [10, 11].

Today, GUIs to produce sound include visual programming languages such as Max/MSP, Pure Data (Pd), OpenMusic (OM), PWGL, Bach, etc. which in some cases already include modules with notation interfaces. Furthermore, there are Trackers, Sequencers, Digital Audio Workstations (DAW) and score editors. Nash et al. continue to research usability dependent on the creative involvement of the user and developing workflow models to address different use-cases by analyzing flow and cognitive dimension metrics [12, 13, 14]. Hunt et al. [15] even uses the cognitive dimensions approach to design an interactive generative score editor from scratch. Nevertheless, detailed academic examination of music notation software remains scarce. Peterson et al. [16] measure duration of various interactions and analyzed their influence on composer creativity. Compared to handwriting they spent less time on the creative task of writing notes due to menu navigation, which resulted in less musical detail.

Taking an analytical view on the components of score editors, they are mostly based on interaction paradigms better known from other contexts: text processing (e.g., inserting notes with the keyboard), image processing (e.g., changing layout with the mouse) and digital audio processing (e.g., placing sound elements in a temporal order and playing them back, similar to DAWs). Although there are convergences in visual and interaction design, a lack of overarching and consensual metaphors leads to more variation between the applications and may create a barrier to change from one to another.

To also account for the use of musically trained people, we decided to approach the problem from two sides—usability and creativity—using established and standardized questionnaires to develop hypotheses about the importance of design elements, which we will be important to explore in further studies.

2. METHOD

Digital musical notation is often aimed at creative use, be it formatting a score beautifully, or for composing or arranging with simultaneous acoustic verification of the result. The interaction is usually conveyed by screen, keyboard

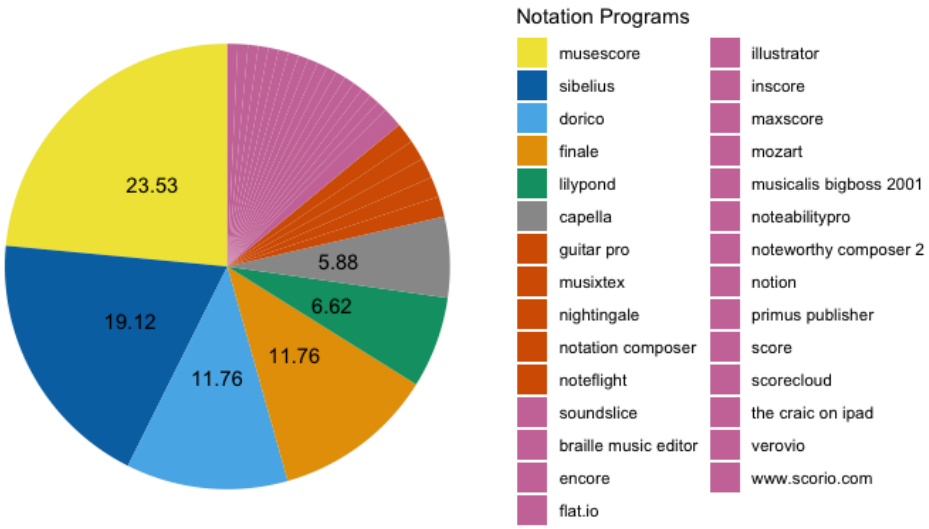

Figure 1. Percentages of all named notation programs in the survey. The dark orange fields equal to 1.47% (named twice) each, while the pink fields equal to 0.74% (named once) each. The first six programs to be analyzed in this paper represent 107 of 136 valid responses.

and mouse, but touch-based solutions are also common. In any case, the design has to ensure good usability and trouble-free use. To cover both aspects, i.e., creativity and usability, we decided to use multiple questionnaires. The System Usability Score (SUS) [17] assesses usability. The AttrakDiff2 [18] assesses pragmatic and hedonistic qualities. The questionnaire by Nash et al. [13] assesses liveness of the interaction.

Pragmatic quality (PQ) measured by the AttrakDiff2 questionnaire is bound to the satisfaction using the software and the feeling of productive impact while using it, whereas hedonistic qualities, which we consider to be desirable goals while working creatively, deal both with:

identity (HQ_I) People express their self through objects. They want to be perceived by relevant others in a specific way. A product can support this by communicating a desired identity.

stimulation (HQ_S) People strive for personal development, i.e., the improvement of knowledge and skills. Products can support this development by having a stimulating effect. Novel, interesting and stimulating functionalities, content, interaction and presentation styles can increase attention, dampen motivation problems or facilitate finding new solutions to existing problems. Thus, stimulation can also indirectly help with task completion.

Considering music production, Nash et al. [13] investigated liveness metrics on trackers and sequencers, which employ different kinds of notation and workflow, whereby score editors can be seen as being similarly manipulation-driven as trackers. Liveness describes generally a sense of subjective sense of intimacy, which can be assessed through system feedback. This is strongly connected to flow, which describes the mental state of immersion while performing a task or using a system [19].

We distributed the survey among mailing lists of interest groups concerning themselves with musical notation in general, digital notation, digital instruments, digital musicology, composition and musical markup languages. [1] By doing this, an international pool of potential participants was addressed. The survey started in July 2022 and lasted until December 2022.

Each participant had to name one editor either by choosing from a preset list or writing into a free text box before answering the questions. Only single mentions of a program were counted as valid entries to ensure differentiated answers. E.g. “musixtex” is valid, while “musixtex and score” is not. The questionnaires could be answered multiple times by the same person, each time answering the questions in relation to another program they have not named before. The participants did not have to be experts with the program but should be confident using it. Further attributes about the participants where not requested. All questions in the survey were mandatory.

We got 138 responses, from which two were not valid. In total, 29 score editors were mentioned. We excluded all editors that were mentioned only once or twice from the further analysis leaving us with six score editors and 107 of the total responses (see Figure 1). The six score editors are: MuseScore, Sibelius, Finale, Dorico, LilyPond and Capella.

MuseScore is available for free, while Sibelius, Finale, Dorico and Capella have to be purchased, but reduced free versions exist. It is worth mentioning that we had nine responses for LilyPond, which is available for free and which follows completely different interaction paradigms than the What-you-see-is-what-you-get (WYSIWYG) in-

[1] The mailing lists were: students at the Detmold University of Music, Music Notation at Ircam (music-notation@listes.ircam.fr), W3 Notation Community Group (public-music-notation@w3.org), NIME Community (nime-community@googlegroups.com), MEI Community (mei-l@lists.uni-paderborn.de)

terfaces we mainly had in mind. It is a notation program, which compiles notations from text files and so is functionally similar to LaTeX.

To make the the questionnaires more comprehensible, we had to add some explanations since concepts condensed in expressions like "viscosity", "diffuseness" or "action-awareness merging", were not always clear in pre-tests in the context of using notation programs. Even after starting the survey we had some remarks about the comprehensibility of some items. Especially for AttrakDiff2 some adjective pairs seemed pointless, as it was communicated to us by English- and German-speaking participants.

Each questionnaire was evaluated as described in their respective papers. To compare the music notation editors, we applied ANOVA on the respective scores for the whole group to determine, if there are any significant differences in the group itself. Additionally η^2 was computed as a metric of confidence about the detected variance. Post-hoc Tukey tests where applied to detect significant differences between pairs of notation editors when ANOVA detected any significant differences. Finally, all metrics were correlated with each other to determine if any metric could be reduced to the outcome of another and what topics might be worth exploring further.

3. RESULTS

3.1 System Usability Score (SUS)

The SUS assesses the general satisfaction of a person dealing with a software. Bangor et al. developed a grading scale that maps SUS scores as follows to school grades [20]:

- below 60: F
- between 60 and 69: D
- between 70 and 79: C
- between 80 and 89: B
- 90 and above: A

Based on 241 studies, the average SUS score is 68 [21, p. 203–204]. Therefore, we added a mark for 68 in Figure 2. Only one of the six programs (Capella) clearly exceeded this threshold. Finale and MuseScore barely exceeded the threshold, while Dorico, Lilypond and Sibelius fall below. Looking at the distribution within the groups we can see large differences. Notation editors, which were mentioned less by our participants (Capella and LilyPond) have smaller ranges between the largest and the smallest scores. Sibelius has an especially wide range.

The one-way ANOVA test shows that there is a significant difference between the classes ($p < .001$). The post-hoc Tukey test shows that there is a significant difference between Capella and all other programs ($p < .01$). Other significant differences are not detected (Table 2).

3.2 AttrakDiff2

The items in this questionnaire consist of polar adjectives, which are ranked on a 7 point Likert scale. A higher number represents a stronger expression of that property.

Overall the one-way ANOVA tests detects significant differences in all three qualities ($p < .001$). Especially HQ_S shows more individual differences and more diverse group parings in the Tukey test (see Table 3). Tukey tests for each quality is characterized by MuseScore and Sibelius having mostly significant differences with most other editors, while Capella, Dorico and LilyPond do not show differences among themselves. HQ_I shows only differences of Capella and Dorico to MuseScore while PQ mirrors the outcome of SUS (see Table 4). In general Capella and MuseScore are considered to be more pragmatic, while Dorico and LilyPond are more stimulating with LilyPond having stronger tendencies towards identity. The scores of Sibelius and Finale are similar to each other across all qualities.

3.3 Liveness

Nash et al. [13, 22] derive their concept of liveness from cognitive dimensions [23] and flow [19] to assess creative work while using programs based on notations. The results in Figure 4 are based on a 5 point Likert scale. Due to the formulations higher values do not always mean a more desirable expression of item. For example "hard mental operations" should be desired to be low, while "no hidden dependencies" is desired to be high, because hidden dependencies may influence prediction of outcomes and may be not controllable by the user. Considering the role of "loss of self consciousness" and "transformation of time" it seems not very clear if lower or higher values are more desirable, but it might be a hint towards more or less rational and controlled handling of the software. However, this did not allow for a meaningful aggregation, so that we evaluated significant differences for every item as demonstrated in Table 5. Due to the relatively high number of examined metrics we decided to choose a significance level of $p < .01$ for one-way ANOVA to reduce random detections of significance. Afterwards Tukey tests were conducted if that significance level was reached. For the Tukey tests, Table 5 reports all group pairings with significance level $p < .05$. For example this tells us, that there are significant differences in "consistency" between Capella & Sibelius and Capella & MuseScore.

It is noticeable that most of the groups in the Tukey test include MuseScore eight times, Capella and LilyPond ar mentioned seven times and Sibelius six times. The most mentioned group is LilyPond & MuseScore with four, followed by Capella & MuseScore, Capella & Sibelus and LilyPond & MuseScore (each three times). LilyPond & Finale and Dorico & MuseScore are mentioned once. MuseScore and Sibelius are often grouped for the same metric (like for "abstraction management" and "intrinsically rewarding") and there are not instances in which significant differences between them can be detected, which is also true for Capella and LilyPond.

Comparing the trajectories of the programs (Figure 5) we can see that Capella is outperforming all others in almost every item. Except for "no premature commitment", "loss of self-consciousness" and "transformation of time" where it is below the level to the best-performing programs. For

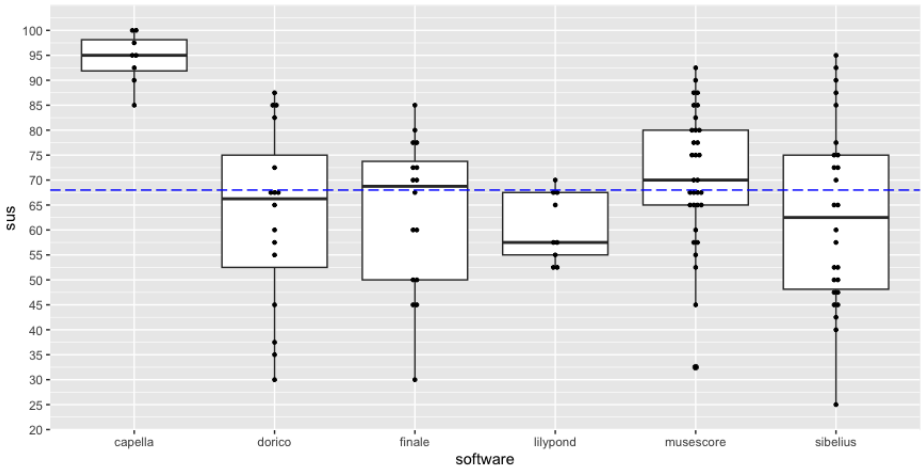

Figure 2. Boxplot of SUS. The points represent the individual results in each group. The dashed line highlights the SUS usability threshold of 68.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η^2
	5	7738.97	1547.79	6.66	2.12e-05	0.57
Residuals	101	23465.82	232.33			

Table 1. ANOVA for SUS

group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
capella	dorico	-31.88	-51.05	-12.70	7.05e-05	****
capella	finale	-31.09	-50.27	-11.92	1.13e-04	***
capella	lilypond	-33.82	-55.34	-12.30	2.00e-04	***
capella	musescore	-23.75	-41.25	-6.25	2.02e-03	**
capella	sibelius	-31.39	-49.30	-13.49	2.38e-05	****

Table 2. Tukey test for SUS. Only groupings with any significance are shown.

"abstraction management" and "virtuosity" LilyPond performs best, which is also reflected by the significance tests.

4. DISCUSSION

We first discuss limitations and problems of our approach. Then we discuss the results in more detail and how they could be interpreted in the context of this study. All metrics were correlated with each other to develop hypotheses about important features of notation software and their interaction design. Since all the results were correlated separately and were not grouped by software, the effect of outlying and biased results is reduced.

4.1 Limitations and Problems

There are several potential problems with our approach in particular and when using an online questionnaire-based approach in general. These limitations and problems are discussed in the following:

Sampling bias: Sampling bias is often seen as a main objection against online surveys as it is difficult or even impossible to achieve a random sample of Internet users [24]. Our approach also suffers from sampling bias: The questionnaire was sent to a large group using mailing lists. The participants were not selected with the help of a systematic or a random process, but decided themselves to participate in the study. Because of this self-selection, it is possible that only those who are specifically interested in the topic have responded, making it difficult to generalize the results to the general population.

Manipulation: While this is not ethical, some survey participants may have deliberately falsified their responses to push their favorite or to harm a competing product as commercial and personal interests may be involved. Furthermore, participants may have submitted the survey multiple times to skew the results or they may hay have encouraged others to rate a score editor in a specific way. While we do not see obvious patterns of manipulation in our dataset, such manipulations can also not be ruled out completely.

Different target groups: Musical notation editors can be used for different tasks ranging from ideation to music engraving. Since each software has a different set of features, some tasks can be completed more efficiently respectively. This has an impact on the

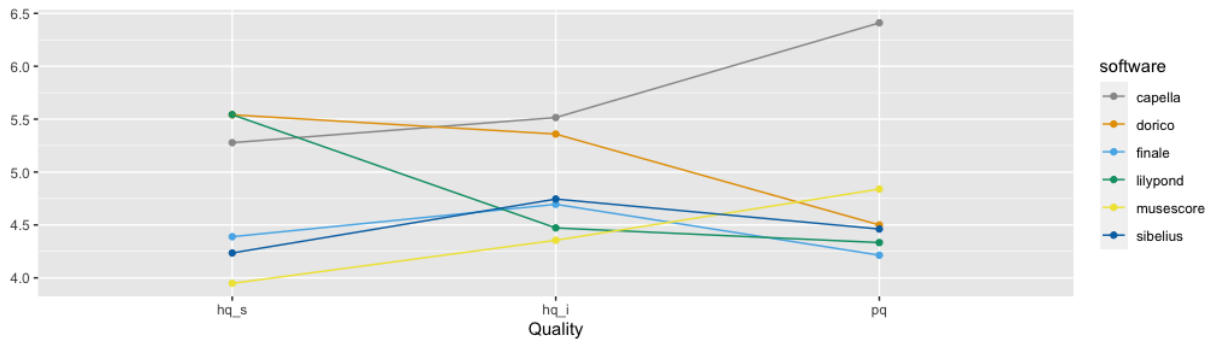

Figure 3. Median values for each quality per software. hq_s = Hedonistisc Quality: Stimulation, hq_i = Hedonistic Quality: Identity, pq = Pragmatic Quality.

Attribute		Df	Sum Sq	Mean Sq	F value	Pr(>F)	η^2
HQ_S		5	62.92	12.58	10.59	3.39e-08	0.72
	Residuals	101	120.00	1.19			
HQ_I		5	23.73	4.75	4.98	0.000405	0.50
	Residuals	101	96.25	0.95			
PQ		5	37.74	7.55	5.32	0.000223	0.51
	Residuals	101	143.37	1.42			

Table 3. ANOVA for AttrakDiff2

Attribute	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
HQ_S	capella	musescore	-1.66	-2.92	-0.41	2.65e-03	**
	capella	sibelius	-1.33	-2.61	-0.05	3.67e-02	*
	dorico	finale	-1.51	-2.63	-0.39	2.12e-03	**
	dorico	musescore	-1.92	-2.89	-0.95	1.40e-06	****
	dorico	sibelius	-1.58	-2.59	-0.58	1.94e-04	***
	finale	lilypond	1.41	0.09	2.73	2.91e-02	*
	lilypond	musescore	-1.81	-3.01	-0.62	3.64e-04	***
	lilypond	sibelius	-1.48	-2.70	-0.25	8.57e-03	**
HQ_I	capella	musescore	-1.51	-2.63	-0.39	0.00228	**
	dorico	musescore	-1.16	-2.02	-0.29	0.00261	**
PQ	capella	dorico	-2.14	-3.64	-0.64	9.79e-04	***
	capella	finale	-2.47	-3.97	-0.97	8.13e-05	****
	capella	lilypond	-2.19	-3.87	-0.51	3.51e-03	**
	capella	musescore	-1.78	-3.14	-0.41	3.62e-03	**
	capella	sibelius	-2.14	-3.54	-0.74	3.17e-04	***

Table 4. Tukey tests for AttrakDiff2. Only groupings with any significance are shown.

target groups which might have more interest in fast note input, or fine grained layout formatting, etc. depending on proficiency and skill level of the user and purpose of the notation.

Long user history: In their responds participants might refer to a longer or shorter history of using a specific software thereby reflecting advantages and shortcomings which occurred over the years, maybe even changing software in this time period.

4.2 Metrics in detail

The ANOVA for SUS shows that there is no significant difference in usability among programs, except compared with Capella. It is surprising that half of the median scores did not even reach the threshold for usability of 68. Two other music notation editors barely exceeded that threshold, which also corresponds to a school mark of “D” as discussed in Section 3.1. However, having a score below the threshold does not mean, that the application is unusable as shown by a study correlating adjectives with SUS. The adjectives ranged from “worst immaginable” to “best immaginable” of which “OK” occupies the space from 50

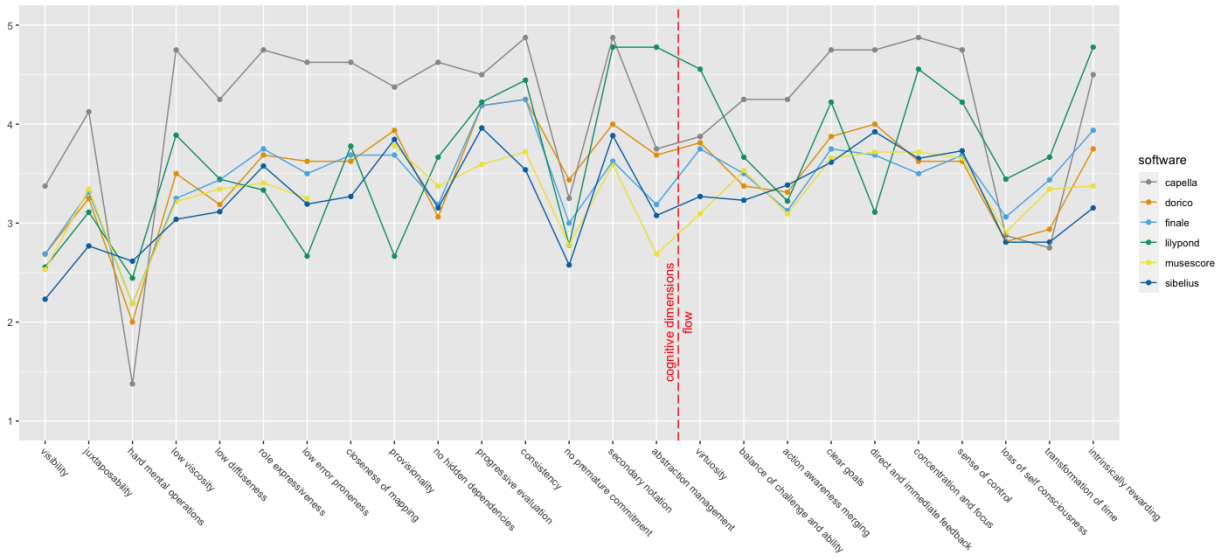

Figure 4. Mean values for each liveness item per software. The dashed line represents the border between items which are derived from cognitive dimensions and flow metrics.

Metric	p-value ANOVA	η^2	Groups Tukey test	p-value Tukey test
Consistency	0.00309	0.44	capella & sibelius capella & musescore	0.0111 0.0355
Secondary notation	0.00513	0.42	capella & musescore lilypond & musescore	0.0250 0.0333
Abstraction management	8.21e-06	0.6	dorico & musescore finale & lilypond lilypond & musescore lilypond & sibelius	2.05e-02 3.68e-03 4.90e-06 4.73e-04
Virtuosity	0.00498	0.42	lilypond & musescore lilypond & sibelius	0.00561 0.02730
Concentration and focus	0.00695	0.41	capella & finale capella & musescore capella & sibelius	0.0239 0.0478 0.0370
Intrinsically rewarding	0.000892	0.47	capella & sibelius lilypond & musescore lilypond & sibelius	0.03620 0.01290 0.00314

Table 5. On the left side, ANOVA tests are shown, if $p < .01$. On the right side, the respective Tukey tests if $p < .05$.

up to 68 [25] and thereby includes lower performing software in our study. It is important mentioning that the SUS was developed for assessing usability of the GUIs specifically. This could explain lower scores for LilyPond in general, since the resulting notation is based on coding text commands. The framing of our survey did not clearly distinguish between a GUI and the process of notating so that when answering the questions the participants relate to the LilyPond language and not the text editor.

There is a noticeable difference in the spread of data points in Figure 2 especially for Sibelius and Dorico, which can be interpreted as heterogenity in the user group. This might be also an effect of the sample size we had in this survey. The values of less mentioned music notation editors such as Capella and LilyPond is usually less spread out. One can assume that different user groups have different technical background. E.g. LilyPond is known to be used by users already familiar with LaTeXor programming. It is not possible to verify such assumptions satisfactorily without qualitative research and knowledge about the circumstances in which the applications are used.

With AttrakDiff2 we wanted to expand the field of interest from pure usability to user experience and emotional responses connected to the application using PQ and SUS as a pivot attribute. In general, we can see that the ranking of the music notation editors matches in both metrics. The exceptional position of Capella in PQ is reflected by significance tests with results similar to SUS (see Table 2 and

Table 4).

Dorico, Capella and LilyPond are considered to be more stimulating than the other programs (see Figure 3). Especially Dorico and LilyPond employ interaction paradigms, which are not found in the other ones. Dorico can be used by opening popovers to create elements. Also notes can be inserted according to beat time divisions within the bar rather than on pre-existing rhythms notated as rests. Furthermore, Note input and layout are strictly separated by different views. The visual output of LilyPond can be completely controlled by manipulating the underlying text files. Future experimental developments might combine fluent transitions between text and GUIs.

Following the individual AttrakDiff2 trajectories for each application (Figure 3), Capella has a similar value for stimulation as Dorico, but overall it is considered more pragmatic than stimulating, which is also true for the trajectory of MuseScore. Finale and Sibelius are balanced over all qualities with slight tendencies towards identity. Noticeable differences from HQ_S to HQ_I is only seen in LilyPond (a drop of 1.1 from 5.5 to 4.4) whereas the remaining programs have differences of between 0.3 to 0.5. Identity and its corresponding items, as described in the original paper can be read in different ways [18]. On the one hand, being perceived by relevant others could be achieved by the software itself through direct communication or collaboration. On the other hand, using it might identify oneself as part of a community. Despite the relative low score, LilyPond is usually used to make scores of higher visual aesthetic quality, by having fine grained control over every visual element. This could be understood as a perceived need for creating shareable scores, contributing to identity.

For Liveness we found six metrics with significant differences, which will may helpful to inform future qualitative research and to examine central design differences. Especially Sibelius and MuseScore are considered to be significantly less consistent than Capella. Secondary notation is the interaction with all graphical objects, which are not the musical notation itself but are rather complementing it, like annotations and coloring. LilyPond and Capella both are rated very high for secondary notation compared to MuseScore. It is however not clear why LilyPond performed that good for secondary notation, since the possibilities of annotating the notation are restricted. However adding secondary notations like comments and formatting the text file is much more straight forward, which might explain the good score. Most significant differences in "abstraction management" are detected by comparison to LilyPond, which shows that knowledge of automated and aggregated actions are valued high for this product although they might be hard to learn. "Virtuosity" is only positively attributed for LilyPond, which is a hint of the value of skillfulness using the program. Capella is significantly different in "concentration and focus" compared to Finale, MuseScore and Sibelius.

Considering the trajectories in Figure 4, we can see again the strong overall performance of Capella. Here we would like to examine some peculiarities in the trajectory of the individual programs and where there are stronger deviations than in others. MuseScore and Sibelius have a very similar profile, while the trajectories of the other programs are much more diverse. For the most of the cognitive dimensions items Sibelius and MuseScore have similar values and are generally underperforming in "consistency", "abstraction management" and "virtuosity" compared to the other programs. LilyPond tends to have lower values with items, which can be explained from it being text-based and the need to be compiled first, like having less feedback and being more prone to errors.

4.3 Correlating the results

By correlating 107 individual results we hope to find important features to investigate closer in future research. We gain a more general view on informative features by *not* grouping the individual results by music notation editor. As correlations that we consider important we decided so set a threshold at $|c| \geq 0.5$ (see Figure 5).

On first sight the negative correlations of "hard mental operations" stand out, which is an artifact because low values are more preferable in this case. SUS and PQ have the highest correlation, which supports our notion that these metrics are connected to the same attributes.

The values of "transformation of time" and "loss of self-consciousness" have mostly weak or no correlations. In turn, we can see more pronounced correlations in action-related metrics, rather than such describing mental states. In general, the matrix represents the connections of our measured and isolated features above by showing high transitive correlations. PQ, "clear goals", SUS and "direct an immediate feedback" are the metrics with the most high ranking values. These are measures of control and they are highly correlated with other measures indicating control over the system. As argued by Csikszentmihalyi [19], an application should support automated behavior, which is based on habit and patterns, and always present its current UI state and what inputs are possible. In a state of flow one is absorbed in the task, and therefore has no resources to reflect on the current action. The outcomes must be based on ordered rules and non-contradictory actions to establish an unbroken experience. Transferred to music notation editors this means that notes and chords should be presented and played directly when selected. Also changing input modalities after inserting a note might lead to break of flow.

"Concentration and focus" represents an important pivot feature, which subsumes correlations with many metrics of flow (such as "clear goals" and "direct an immediate feedback") and cognitive dimensions alike. It is also correlated with usability metrics, but it is mostly associated with pragmatic, rather than hedonistic qualities. This is also true, e.g., for "role-expressiveness", which expresses whether the user can see how each component of a program relates to the whole [26, 23]: The purpose and condition of the musical structure should therefore be readily visible and the relationships should be easy to see. This can be achieved by overview or analysis functions to make see harmonic relationships (e.g. by proposing chord names) or by having a good view over all instruments while mak-

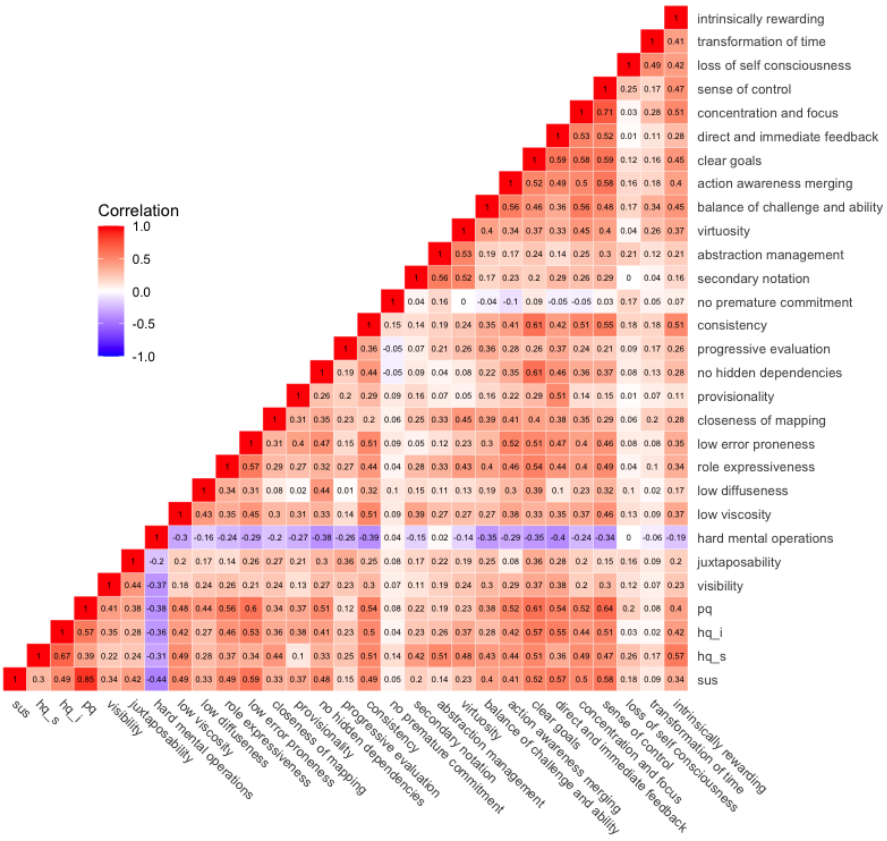

Figure 5. Correlation matrix of all individual results.

ing changes thus pointing to qualities of changing between views of the same score and reduction of distracting and irrelevant elements.

In Table 5 "Virtuosity" has mostly shown differences for LilyPond, but in general it is mostly associated with "secondary notation" and "abstraction management". "Virtuosity" is a widely used term in music and comparing it to HQ_S they both reflect skillfulness. When designing software that helps to develop skills in certain fields, be it composition, musicology, editing or music practice, one might need to identify more detailed or overlapping goals. This is not necessarily bound to musical skills, but also to the learnability of the system itself and knowing the tools and possibilities to create a score with creative needs in mind.

In summary, we consider the following metrics as the most informative for music score editors:

- SUS
- HQ_S
- Consistency
- Secondary notation
- Abstraction management
- Virtuosity
- Concentration and Focus
- Role expressiveness

5. CONCLUSION & FUTURE WORK

In this paper we examined usability and user experience of music notation editors with the help of an online survey with standardized questionnaires. Of course there are limitations to such an approach (sampling bias, manipulation, different target groups and long user history), but this study provides a first starting point for a scientific examination of existing music notation editors. Almost all examined music notation editors show weak results in usability as measured by the SUS. The exception is Capella that achieved a school grade of "A" and an adjective rating of "Excellent" according to the rating scheme of Bangor et al. [25].

User experience questionnaires, like AttrakDiff2 [18] and the liveness questionnaire by Nash et al. [13], examine further important aspects of working with music notation editors. While the pragmatic quality of the AttrakDiff2 was highly correlated with SUS (as expected), the hedonistic qualities stimulation (HQ_S) and identity (HQ_I) were strongly differentiating factors between the programs. It seems that non-standard interaction paradigms such as text input can lead to higher values in both hedonistic qualities. Comparing results for the AttrakDiff2 and liveness metrics, MuseScore and Sibelius are very similar.

We correlated all individual results. Metrics with implications for action and control correlated relatively strongly with each other as opposed to metrics that fit better in creative contexts.

Our study of usability and user experience of music nota-

tion editors, can help to create experimental music notation programs. Mixtures of different interaction paradigms, like those used in Dorico and LilyPond, could be developed. Future music notation editors could even be adapted by gradually moving from one paradigm to the other. We also expect interesting insights from cognitive and eye-tracking studies. This could be beneficial to assess specific interaction patterns, also in comparison to non-digital music writing [16].

6. REFERENCES

[1] N. Bevana, J. Kirakowski, and J. Maissela, "What is Usability?" in *Proceedings of the 4th International Conference on HCI*, 1991, pp. 1–6.

[2] T. L. Roberts and T. P. Moran, "The evaluation of text editors: methodology and empirical results," *Communications of the ACM*, vol. 26, no. 4, pp. 265–283, 1983. [Online]. Available: https://doi.org/10.1145/2163.2164

[3] N. S. Borenstein, "The evaluation of text editors: a critical review of the Roberts and Morgan methodology based on new experiments," *ACM SIGCHI Bulletin*, vol. 16, no. 4, pp. 99–105, 1985. [Online]. Available: https://doi.org/10.1145/1165385.317475

[4] S. K. Card, T. P. Moran, and A. Newell, "The keystroke-level model for user performance time with interactive systems," *Communications of the ACM*, vol. 23, no. 7, pp. 396–410, 1980.

[5] L. A. Hiller and R. A. Baker, "Automated music printing," *Journal of Music Theory*, vol. 9, no. 1, pp. 129–152, 1965. [Online]. Available: http://www.jstor.org/stable/843151

[6] D. A. Byrd, "Music notation by computer," Ph.D. dissertation, Indiana University, 1984.

[7] G. Loy and C. Abbott, "Programming languages for computer music synthesis, performance, and composition," *ACM Computing Surveys*, vol. 17, no. 2, p. 235–265, 1985. [Online]. Available: https://doi.org/10.1145/4468.4485

[8] R. B. Dannenberg, "Music representation issues, techniques, and systems," *Computer Music Journal*, vol. 17, no. 3, pp. 20–30, 1993. [Online]. Available: http://www.jstor.org/stable/3680940

[9] D. V. Oppenheim, "The Need for Essential Improvements in the Machine-Composer Interface used for the Composition of Electroacoustic Computer Music," in *Proceedings of the International Computer Music Conference (ICMC)*, 1986.

[10] ——, "The P-G-G Environment for Music Composition: A Proposal," in *Proceedings of the 1987 International Computer Music Conference (ICMC)*, 1987, pp. 40–48.

[11] M. S. Puckette, "The Patcher," in *Proceedings of the International Computer Music Conference (ICMC)*, 1988.

[12] C. Nash and A. Blackwell, "Tracking Virtuosity and Flow in Computer Music," in *Proceedings of the International Computer Music Conference (ICMC)*, 2011.

[13] ——, "Liveness and Flow in Notation Use," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2012. [Online]. Available: http://www.nime.org/proceedings/2012/nime2012_217.pdf

[14] ——, "Flow of Creative Interaction with Digital Music Notations," in *The Oxford Handbook of Interactive Audio*, H. T. Kasey Collins, Bill Kapralos, Ed. Oxford University Press, 2014. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780199797226.013.023

[15] S. Hunt, T. Mitchell, and C. Nash, "A Cognitive Dimensions Approach for the Design of an Interactive Generative Score Editor," in *Proceedings of the International Conference on Technologies for Music Notation and Representation (TENOR'18)*, 2018, pp. 119–127.

[16] J. Peterson and E. Schubert, "Music Notation Software: Some Observations on its Effects on Composer Creativity," in *Proceedings of Intercational Conference on Music Communication Science (ICoMCS)*, vol. 127–130, 2007.

[17] J. Brooke, "SUS: A Quick and Dirty Usability Scale," *Usability Evaluation in Industry*, vol. 189, no. 194, pp. 4–7, 1996.

[18] M. Hassenzahl, M. Burmester, and F. Koller, "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität," in *Mensch Computer 2003: Interaktion in Bewegung*, 2003, pp. 187–196.

[19] M. Csikszentmihalyi, "Flow and the psychology of discovery and invention," *Harper Perennial*, vol. 39, 1997.

[20] A. Bangor, P. Kortum, and J. Miller, "The system usability scale (SUS): An empirical evaluation," *International Journal of Human-Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[21] J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[22] C. Nash, "The Cognitive Dimensions of Music Notations," in *Proceedings of the First International Conference on Technologies for Music Notation and Representation (TENOR'15)*, 2015, pp. 190–202.

[23] T. R. G. Green and M. Petre, "Usability analysis of visual programming environments: A 'cognitive dimensions' framework," *Journal of Visual Languages & Computing*, vol. 7, no. 2, pp. 131–174, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1045926X96900099

[24] M. Van Selm and N. W. Jankowski, "Conducting online surveys," *Quality and quantity*, vol. 40, no. 3, pp. 435–456, 2006.

[25] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *J. Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.

[26] A. Blackwell, C. Britton, A. Cox, T. Green, C. Gurr, G. Kadoda, M. Kutar, M. Loomes, C. Nehaniv, M. Petre, C. Roast, C. Roe, A. Wong, and R. Young, "Cognitive Dimensions of Notations: Design Tools for Cognitive Technology," in *Cognitive Technology: Instruments of Mind*, M. Beynon, C. L. Nehaniv, and K. Dautenhahn, Eds. Springer-Verlag, 2001, pp. 325–341.