MIREX GRAND CHALLENGE 2014 USER EXPERIENCE: QUALITATIVE ANALYSIS OF USER FEEDBACK

Jin Ha Lee
University of Washington
jinhalee@uw.edu

Xiao Hu
University of Hong Kong
xiaoxhu@hku.hk

Kahyun Choi, J. Stephen Downie
University of Illinois
{ckahyu2,jdownie}@illinois.edu

ABSTRACT

Evaluation has always been fundamental to the Music Information Retrieval (MIR) community, as evidenced by the popularity of the Music Information Retrieval Evaluation eXchange (MIREX). However, prior MIREX tasks have primarily focused on testing specialized MIR algorithms that sit on the back end of systems. Not until the Grand Challenge 2014 User Experience (GC14UX) task had the users' overall interaction and experience with complete systems been formally evaluated. Three systems were evaluated based on five criteria. This paper reports the results of GC14UX, with a special focus on the qualitative analysis of 99 free text responses collected from evaluators. The analysis revealed additional user opinions, not fully captured by score ratings on the given criteria, and demonstrated the challenge of evaluating a variety of systems with different user goals. We conclude with a discussion on the implications of findings and recommendations for future UX evaluation tasks, including adding new criteria: Aesthetics, Performance, and Utility.

1. INTRODUCTION

Since 2005, the Music Information Retrieval (MIR) community has benefited from the Music Information Retrieval Evaluation eXchange (MIREX), an annual evaluation event led by researchers at University of Illinois [7]. MIREX has had a significant contribution to the field as it allows system developers to test and improve their MIR algorithms. However, as the field matures, the current state of the art is increasingly deemed sufficient to support an acceptable degree of efficiency and effectiveness in various conventional MIREX tasks, resulting in the glass ceiling effect [1,2,11]. A number of researchers have also pointed out the limitations of MIREX, including the dominance of a system-centered approach and the lack of consideration for real users [11,12,14,19].

In response to the feedback received from the MIR community, the MIREX grand challenge was held in 2014. This was substantially different from any of the past evaluation tasks in two respects: 1) the focus of evaluation shifted to include the front end of the system (i.e., how users interact with the system), and 2) the submissions were complete MIR systems that can employ various MIR techniques rather than individual algorithms. This marks a shift of the evaluation paradigm, since all the MIREX evaluation tasks have been focused on the back end, with the front end being largely ignored [11,15].

Three different MIR systems participated in the Grand Challenge 2014 User Experience (GC14UX). In this paper, we present the findings from analyzing the results of GC14UX, focusing on the free-text user responses. The goal of the paper is twofold: 1) understanding how users reacted to which aspects of the systems in their responses, and 2) using that knowledge to improve the design of future MIR UX evaluation tasks. In particular, we seek to answer the following research questions:

Q1. Which aspects of MIR systems were most important to users, as evidenced by the responses?

Q2. Based on users' responses, are there any evaluation criteria we should consider revising or adding for future iterations of MIR evaluation of user experience?

2. BACKGROUND

2.1 User-centered Evaluation in MIR

As pioneers in user-centered evaluation in MIR, Pauws et al. [17,18,20] conducted a series of user evaluation tasks to examine an interactive playlist generation system. Several user-centered measures were considered, including time on tasks, number of actions, preference, ease of use, and usefulness. Although the evaluation was confined to one specific MIR system, it is noteworthy that they considered the front-end interface and the user's interaction in the earlier days of MIR system evaluation. Hoashi et al. [9] also conducted a user evaluation of visualization interfaces for MIR systems based on subjective measures such as perceived accuracy and enjoyability.

Despite such efforts, most MIR evaluation research is still based on a system-centered approach without involving users. While this makes sense for some of the micro-level tasks, ultimately many algorithms that are being evaluated will be implemented as features in complete MIR systems. Therefore it is important to consider how users determine the usefulness and value of the systems. Hu and Kando [10] also emphasized the need for user-centered evaluation in MIR based on their finding that only a weak correlation existed between user-centered measures and system-centered measures in their evaluation experiment of MIR systems.

Leaving aside the shortage of user-centered evaluation in our field, the evaluation in the few aforementioned studies has been mostly limited to specific algorithms or

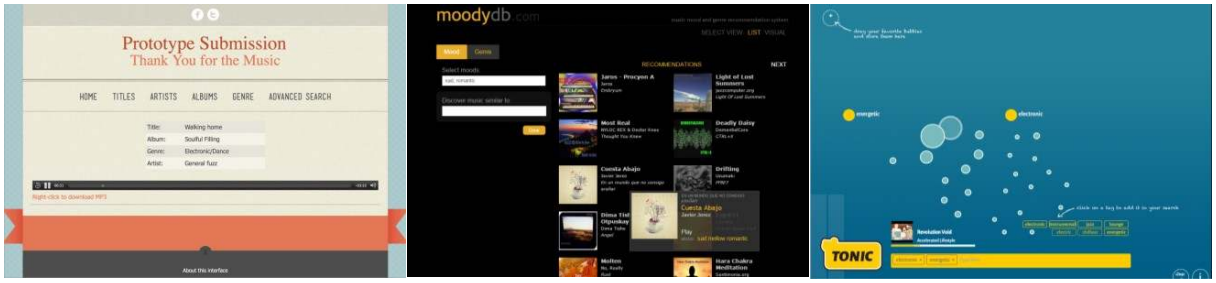

Figure 1. Screenshots of *Thank You for the Music*, *Moody*, and *Tonic.*

functions such as playlist generation algorithms and recommender systems [14]. This has been attributed to the lack of complete MIR systems ready for evaluation within the MIREX framework [11,15]. Consequently, attempts to conduct a holistic user-centered evaluation of MIR systems had to be done with existing commercial music services. For example, Lee and Price [15] examined how Nielsen's usability heuristics [16] can be applied to evaluate multiple aspects of user experience for services like Pandora, Spotify, etc. As the MIR field is maturing, there is also a growing recognition that we are ready to evaluate complete and full-featured systems incorporating various sub-components with helpful interfaces [6,11]. Therefore, GC14UX was held, aiming to inspire the development of complete MIR systems and a holistic evaluation of user experience with those systems.

2.2 GC14UX Evaluation Framework and Process[1]

The dataset used in GC14UX was a sample of 10,000 tracks with the CC-BY (Creative Commons Attribution) license from the Jamendo collection[2], for the purpose of avoiding any potential copyright issues. All tracks had song and album titles, artist name, and at least two genre tags. To guide the evaluators, a user task was created based on several criteria: 1) a common and realistic MIR task, 2) a task specific enough to help evaluators judge how successful the results are, 3) a task not tied to a particular MIR technique, and 4) a task that can be reasonably accomplished with the given dataset. The final task was determined as follows: "You are creating a short video about a memorable occasion that happened to you recently, and you need to find some (copyright-free) songs to use as background music."

An online evaluation platform was set up so that evaluators could easily access the MIR systems through a web browser. The invitations were circulated through mailing lists within the MIR community. Evaluators were asked to interact with the systems and rate their scores on a seven-point Likert scale for the following criteria:

- **Overall Satisfaction**: How would you rate your overall satisfaction with the system?
- **Learnability**: How easy was it to figure out how to use the system?
- **Robustness**: How good is the system's ability to warn you when you are about to make a mistake, allow you to recover, or retrace your steps?
- **Affordance**: How well does the system allow you to perform what you want to do?
- **Feedback**: How well does the system communicate what is going on?

Evaluators were also given an opportunity to provide their comments in an open text field.

2.3 Participating Systems and Quantitative Ratings

There were a total of three systems that participated in GC14UX: *Thank You for the Music* (hereinafter, *Thank You*), *Moody,* and *Tonic* (Figure 1)[3]. The design and functionality of the three MIR systems varied to some extent. *Thank You* provides users with access to a music collection through a more traditional music digital library interface, offering music search by title, album, genre, and artist. *Moody* is a recommender system in which a music collection can be browsed based on mood and genre. *Tonic* is a tag-based discovery system with a highly interactive interface utilizing pre-defined tags to find songs.

The three systems received mean scores between 4.15 and 5.37 across all criteria [11]. *Tonic* received the best score in Affordance (4.71), Feedback (4.79), and Overall Satisfaction (OS) (5.11). *Thank You* scored the highest in Learnability (5.37) with an OS of 4.15. *Moody* led in Robustness (4.53) with an OS of 4.63. However, the results of the Kruskal-Wallis test [5] showed that only the OS category had significant differences across systems [11].

3. ANALYSIS OF USER FEEDBACK

3.1 Codebook and Coding Process

We employed content analysis, a widely used qualitative data analysis method as described in [13], to uncover and code common themes in the 99 user responses. On average, there were 69 words in a response (median=51, max=259, min=2). The codebook was developed through an iterative process involving test-coding a subset of data and revising the codes for clarity. Table 1 presents detailed information on all the codes that emerged from the user responses. Each user response contained an average of 3.17 excerpts, each representing a particular code. The codes were organized into seven higher-level categories based on topical similarity. The count of excerpts for each code and the percentage calculated over the total number of excerpts (314) are also reported in the table.

[1] For more detailed information on the framework, see [11].
[2] https://www.jamendo.com/en/welcome
[3] Accessible at: http://bit.ly/1zqz1m0 (*Thank You*), http://bit.ly/1R3rNdr (*Moody*), http://bit.ly/1GU7GLO (*Tonic*)

	Categories	Codes	Definition	#	%
Evaluation Criteria	Aesthetics	attractiveness	The user specifically talks about the visual appeal of the interface.	27	8.6
	Affordance	access	The user specifically comments on an ability to access original music files within the system.	7	2.2
		play function	The user specifically talks about the music play function in the system, including various aspects of the player such as the interface and features.	25	8.0
		save function	The user specifically talks about some kind of save function like a bookmark function allowing users to revisit the page, ability to save the selected songs, or preservation of specific system settings set by the user.	12	3.8
		search/ browse	The user specifically mentions topics related to searching or browsing music based on metadata (e.g., artist name, song/album title, genre, mood labels), including advanced search, auto-complete, and finding similar items.	91	29.0
	Feedback	clarity	The user specifically talks about the clarity of functions or labels provided.	39	12.4
	Learnability	ease of use	The user talks about how easy, intuitive, and user-friendly it is to use the system and complete their desired task.	40	12.7
		help	The user comments on help provided in the system such as guidelines, tutorials, or instructions.	9	2.9
	Performance	bugs/ glitches	The user specifically talks about bugs/glitches in the system that cause it to produce incorrect or unexpected results, or behave in unintended ways.	5	1.6
		response time	The user specifically talks about the response time (i.e., the length of time taken for a system to react to a given event).	8	2.5
		search results	The user specifically talks about the quality of search results and how they are presented to the user.	32	10.2
	Utility	usefulness	The user talks about the overall usefulness of the system, as well as its usefulness for the given evaluation task.	13	4.1
Additional aspects	External factor	dataset	The user specifically notes the effects and/or limitations of using a particular dataset for the evaluation task.	6	1.9
	Sentiment	positive	The user expresses positive feelings in terms of a particular code.	107	34.1
		negative	The user expresses negative feelings or desires for specific functions/features in terms of a particular code.	198	63.1

Table 1. Summary of codebook.

The first six categories correspond to particular evaluation criteria. We can observe that three of these categories were used as evaluation criteria in GC14UX (in bold). Codes matching the criterion Robustness did not emerge from coding user responses. The External factor category contains the code dataset that was used to mark the responses noting limitations of the experience due to variables that were not controllable by system developers. We also had an "Other" code used for uncommon but relevant part of responses that did not fit into existing codes (e.g., comments on scalability issues, mobile device compatibility, etc.). Codes in the Sentiment category (i.e., positive and negative) were used in conjunction with another code to note users' feelings regarding that code.

3.2 Inter-coder Reliability

To ensure consistent application of codes, two coders were recruited. The coders independently coded a subset of user excerpts (42% of all excerpts) and Cohen's kappa coefficient [4] was calculated to measure their agreement. Table 2 shows that all the kappa coefficients for each code fall in the range of good (.60-.74) or excellent agreement (.75-1.0) [3,8]. The Pooled Kappa statistic summarizing the overall results across all the codes [21] was .884, suggesting an excellent agreement.

Code	Kappa value	Agreement level
save function	1.00	excellent
bugs/glitches	1.00	excellent
negative	0.98	excellent
positive	0.97	excellent
play function	0.95	excellent
response time	0.92	excellent
help	0.91	excellent
dataset	0.88	excellent
attractiveness	0.87	excellent
clarity	0.86	excellent
usefulness	0.85	excellent
ease of use	0.82	excellent
search/browse/metadata	0.80	excellent
search results	0.80	excellent
access	0.66	good

Table 2. Kappa coefficients for each code.

3.3 Tabulation of Codes

Table 3 shows the counts of positive excerpts for each system, sorted by the sum of all counts for each code. We can observe that participants liked *Thank You* for more functional reasons (e.g., search/browse, access to music files, search results) whereas they liked *Tonic* for aesthetics and usability aspects (e.g., attractiveness, ease of use, usefulness) in addition to functional reasons (e.g., play function, save function). *Moody*'s scores were fair across most of the codes except save function, access to music files, and search results. Overall, *Tonic* had the highest number of positive excerpts, with *Thank You* and *Moody* having approximately the same numbers.

	Thank You	*Moody*	*Tonic*	Sum
search/browse	**14**	10	4	28
ease of use	8	9	**10**	27
attractiveness	0	7	**11**	18
usefulness	1	1	**6**	8
play function	1	2	**3**	6
save function	0	0	**6**	6
access to music files	**4**	0	0	4
clarity	1	1	1	3
help	0	1	**2**	3
response time	1	1	1	3
search results	**1**	0	0	1
Total	31	32	44	107

Table 3. Tabulation of positive codes.

We also tallied up the counts of negative excerpts for each system (Table 4). Negative excerpts also include desires for additional features/functions, so a high count does not necessarily mean that participants disliked the system. *Moody* had the highest number of negative excerpts, mostly for search/browse, which was also the most commonly mentioned aspect across all three systems. Evaluators had strong opinions about the search function in *Moody*, also evidenced by the highest number of counts in search results. For *Tonic*, improving the clarity and help was important, in addition to play function.

	Thank You	*Moody*	*Tonic*	Sum
search/browse	19	**34**	10	63
clarity	5	10	**20**	35
search results	3	**11**	9	23
play function	3	7	**9**	19
ease of use	4	2	**7**	13
attractiveness	**4**	**4**	1	9
save function	2	**4**	0	6
dataset	2	**3**	1	6
help	0	1	**5**	6
bugs/glitches	**2**	1	**2**	5
response time	**3**	1	1	5
usefulness	2	**3**	0	5
access to music files	0	1	**2**	3
Total	49	82	67	198

Table 4. Tabulation of negative codes.

When we tabulate the counts based on the top-level categories and compare the counts for positive and negative excerpts for each category, we can observe with which aspects evaluators were most satisfied and dissatisfied (Table 5). Across all three systems, Affordance, Performance, and Feedback had more negative excerpts, suggesting these aspects need to be improved upon. Learnability, Aesthetics, and Utility had more positive excerpts overall, although notably *Thank You* had no positive excerpt for Aesthetics.

Top level	*Thank You*	*Moody*	*Tonic*	Sum
Affordance +	19	12	13	44
Affordance -	**24**	**46**	**21**	**91**
Learnability +	**8**	**10**	12	**30**
Learnability -	4	3	12	19
Feedback +	1	1	1	3
Feedback -	**5**	**10**	**20**	**35**
Performance +	2	1	1	4
Performance -	**8**	**13**	**12**	**33**
Aesthetics +	0	**7**	**11**	**18**
Aesthetics -	**4**	4	1	9
Utility +	1	1	**6**	**8**
Utility -	**2**	**3**	0	5

Table 5. Tabulation of codes at the top level categories.

4. DISCUSSION OF CATEGORIES AND CODES

4.1 Aesthetics

Aesthetics consists of a single code regarding the overall attractiveness of the system. While this aspect was not included in the GC14UX evaluation criteria, it may be appropriate to consider adopting it for future iterations. Most excerpts coded with attractiveness were about how appealing the visual interface was, with a few comments about the use of white space, clean interface, use of animation, and background color. The importance of this aspect is well-captured in the following response:

> *"What's funny is that while [Thank You] allows me to search and browse, I really liked the graphic nature of the previous two interfaces. I don't necessarily think this interface performs any less well than the others--"*

4.2 Affordance

Affordance consists of four codes related to particular features or functions of the system. For access, most of the excerpts mentioned that *Thank You* was the only system where users could download the songs. To some participants, that meant that the system was "complete," and gave them "real results."

Excerpts coded with play function tended to be more negative, mentioning evaluators' desires to have more control in which part of the song they are playing. Some evaluators did appreciate that *Tonic* plays the selected songs starting in the middle (e.g., *"I like the fact that the selected pieces start playing from the middle, giving an immediate sense of the general mood and texture of the piece"*), but more evaluators wanted to be able to select from multiple options themselves:

> *"I would like to have an [sic] checkable option, "start playing from beginning"/"start playing from the middle" (or 25%, 30%, 40%), because sometimes [what] is important [is] the beginning, and sometimes (mostly) the mood of whole song."*

Evaluators also commented negatively on the fact that they had to go through another step for playing the music in *Thank You* and *Moody* (e.g., *"I'd expected to start playing a track whenever I clicked on its cover, instead of having to wait for the pop-up and click 'play.' "*). The lack of visibility of the play button/slider was also noted for *Moody* and *Tonic* (e.g., *"The 'play'-slider is a bit small"*; *"difficult to find play button for the next song"*).

With regard to the save function, *Tonic* had multiple positive excerpts on the usefulness of the bookmark function, which was missing in other systems:

> *"there is a function at the top left corner for users to save their favorite results, it is convenient for user to compare the music later and choose the best result…" [Tonic]*
> *"i wish there were a way for me to create[,] like a list, collection, playlist, or save or favorite multiple songs for comparison or reconsideration." [Moody]*

The system remembering user settings was also important; evaluators noted that in *Thank You*, the player does not keep the selected volume level when a new song is loaded, and in *Moody*, switching between the mood and genre tab discards the selected search parameters. The save function code is somewhat related to the Robustness criterion in GC14UX; users want features that will help them trace back and return to previous results, although no excerpts were related to recovering from an error.

Overall, the search/browse code was applied most often, excluding the sentiment codes. *Thank You* had the highest number of positive excerpts due to the fact that multiple search options were provided (i.e., text search, form search, and advanced search) and users had the most control over how the search could be conducted (e.g., *"Its searching technique is very comprehensive and fully developed, which is excellent for users to carry out detailed and accurate search"*). The auto-complete features in *Moody* and *Tonic* were also appreciated by multiple evaluators. However, there was still a lot to be desired from the search/browse functions in all three systems. For *Thank You*, the lack of a browsing mechanism and inability to get recommendations were noted. Evaluators also commented on the limitation of genre categorization:

> *"...about 255 songs are identified as unknown, it may cause inconvenience to the users as they do not know the type of song, they must spend time to listen [to] it first."*

For *Moody*, nine excerpts specifically asked for an ability to combine both mood and genre for search. Some wanted more labels for mood and genre, and others noted the lack of a free-text search option. For *Tonic*, a few evaluators commented on the inaccuracy of certain labels and a lack of vocabulary control:

> *"…the connection from tags to audio content does not always seem to be 'correct'…Especially if more than two tags are combined, there seem to be some problems."*
> *"Moreover, maybe due to the vocabulary control, when I type 'cheerful,' no result is found, I have to type 'happy' instead, so the system is not flexible enough."*

4.3 Feedback

This category consists of a code "clarity" that is about how intuitive and clear the functions and labels were. *Tonic* had 20 negative excerpts that were primarily about evaluators having trouble understanding what information the different design constructs are trying to convey (e.g., meaning of the histogram, size of the bubble) or what the result of a particular user action was:

> *"For instance, I noticed some bars on the left side, each corresponding to one of the search terms, that varied in height along with the bubble, which also resized. What is that? What does it mean when I move bubbles around?"*

Similar concern was also raised for *Moody* (e.g., *"what does the size of the image mean?"*). In addition, a couple of evaluators pointed out that they had a hard time figuring out what the "discover music similar to" function was supposed to do. For *Thank You*, several evaluators commented on misunderstanding genre ID as the count of items under a particular category.

4.4 Learnability

This category contained two codes: ease of use and help. Overall, there were a lot more positive than negative excerpts regarding the ease of use across all three systems. Simple, intuitive, and user-friendly interface design was appreciated for *Moody* and *Tonic*. In general, evaluators also found the basic search interface in *Thank You* easy to use. Negative excerpts were on issues like the page layout or too much text (*Thank You*) or opinions based on a comparison with other systems (e.g., *"Tonic is not that easy to use when comparing with Moody"*).

For the help code, evaluators commented positively on the usefulness of a short introduction on how to use the system for *Tonic* but still desired more explanation on the meaning of design elements. For *Thank You* and *Moody*, clear searching guidelines and limitations (e.g., *"Maybe it should say somewhere that the similarity search only works for artists in the database"*) were desired.

4.5 Performance

Of the three codes belonging to the Performance category (i.e., bugs/glitches, response time, search results), search results was most commonly used, and primarily with negative sentiment. For *Thank You*, the ability to sort the results was appreciated but different sorting criteria were desired. The lack of a sorting mechanism was also mentioned for *Moody*. In addition, three evaluators stated that they wanted to know how many results there are for a particular search, as well as an option to switch between AND and OR connectors. For *Moody* and *Tonic*, several evaluators commented that they did not agree with or could not understand the results:

> *"The returned music doesn't really fit the moods, especially 'romantic.' " [Moody]*
> *"I wrote: 'piano' and 'jazzy,' and just in the middle between these two main bubbles I found the song 'Salmacis – Arkangel,' which is not piano nor jazzy at all." [Tonic]*

The evaluators' reactions to response time tended to vary, even for the same system, possibly due to varying Internet connection speeds and different levels of expectation. Bugs and glitches in scrolling, music playback, and entering data were also mentioned a few times, but they could also depend on the resolution setting or other configurations of the evaluators' machines and browsers. Therefore, it is important to note that what we are seeing

is simply users' interpretation of how well the system performed rather than the objective performance level.

4.6 Utility

"Usefulness" was the only code in this category, noting the general usefulness of the system as well as its appropriateness for the specified user task. *Tonic* had all positive excerpts as evaluators deemed that the tag-based browsing interface worked well for unknown music. The negative excerpts on *Thank You* and *Moody* mostly showed that evaluators wanted more features and functions. For *Thank You*, one evaluator noted that the search interface is limiting for the given evaluation task, which is about finding music for editing a personal video, since there are no content-based features.

4.7 External Factor

Comments on the limitation of the dataset were captured using the code dataset in this category. Six excerpts marked with this code were all negative, mostly stating that evaluators' unfamiliarity with the songs hindered their ability to effectively use the systems. This was especially true for *Thank You*, as evaluators could not issue searches using metadata such as artist name or song title. One evaluator also noted the difficulty in ascertaining the cause of unsuccessful results:

> *"…maybe the Jamendo collection is not very good for the task because of its variability: do we really not have good results or are systems unable to find them?"*

5. IMPLICATIONS ON UX EVALUATION IN MIR

Based on the user responses and the experience of running GC14UX, we discuss three main implications for future UX evaluation tasks in MIR:

1) Adjustment of evaluation criteria

We recommend considering new criteria, Aesthetics, Performance, and Utility, in future UX evaluation tasks. The quantitative ratings showed that the difference of the scores in the "Overall Satisfaction" was statistically significant, but the differences in the other four criteria were not. This suggests that perhaps there are additional evaluation criteria affecting users' overall satisfaction. Based on the responses, the visual aesthetics of the system seem especially important; it is noteworthy that a large proportion of positive excerpts for *Tonic*, the most highly rated system, were based on "Aesthetics". "Aesthetics" might be the missing piece that can explain the differences observed in the "Overall Satisfaction". We also recommend rethinking the criterion "Robustness"; this may be difficult to evaluate given the limited time evaluators have to interact with the systems in the MIREX framework.

2) A better dataset and more user tasks

As some users pointed out, lack of familiarity with the songs in the dataset hindered their search/browse experience. In addition, a single user task for evaluation seems limiting, as MIR systems can serve a wide variety of use cases and scenarios. This was in fact the case in GC14UX as the three evaluated systems were designed to serve different goals (e.g., *Thank You* for known-item searches, *Moody* for mood and genre-based search/browsing, and *Tonic* for exploring new music based on tags). For future UX evaluation, it might be worthwhile to consider establishing multiple user tasks, and perhaps something more common (e.g., playlist generation, recommendation) rather than trying to creating a task suitable for the dataset.

3) Focus on evaluation rather than competition

In addition to a common user task for evaluation, it may be fruitful to consider asking system developers to define a user task for which they want their system to be evaluated, as a secondary task. This makes sense considering that many commercial MIR systems are often targeted to support specific MIR tasks (e.g., Pandora for online radio function, Shazam for music identification), which was also the case for the three evaluated systems from GC14UX. We do acknowledge that this means we will not be able to directly compare the evaluation results of multiple systems. However, we strongly believe that the community should move away from considering this evaluation as a competition where ranking the systems is the primary goal. If we treat this as an opportunity to evaluate the systems in order to improve the design of all participating systems rather than being able to claim one system is better than the other, this issue will naturally dissolve. In case of GC14UX, the differences in scores for the three systems were not substantial; even for the single category where there was a statistically significant difference among the scores (i.e., Overall Satisfaction), the difference between the best- and the worst-performing systems is less than one point in a seven-point Likert scale (5.11 vs. 4.15). What would truly benefit our community as a whole is learning from the feedback about what users need and want, which will inform us on how to improve the design of MIR systems in general.

6. CONCLUSION AND FUTURE WORK

GC14UX was the very first attempt in conducting a holistic evaluation of user experience for complete MIR systems in the history of MIREX. Therefore, reflecting on our experience and deliberating on how to improve future UX evaluation is critical. Our findings indicate which aspects of the systems most concerned users, and how we can use that knowledge to improve the design of and criteria for future UX evaluation. We discussed three key implications for future UX evaluation: 1) consider three new criteria in future UX evaluation tasks, 2) seek a better dataset to improve evaluators' ability to effectively use the features and judge the quality of the results, and select more user tasks to reflect the diversity of the systems, and 3) focus on evaluation for the improvement of systems rather than competition. We hope to continue UX evaluation as a regular task within MIREX, and redesign the task with new use scenarios and datasets in the future. We also plan to widen our pool of evaluators so that we can do a comparative analysis of how MIR experts and general users evaluate their experiences.

7. REFERENCES

[1] J.-J. Aucouturier and F. Pachet: "Improving timbre similarity: how high's the sky?" *Journal of Negative Results in Speech and Audio Sciences*, Vol. 1, No. 1, pp. 1-13, 2004.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri: "Automatic music transcription: breaking the glass ceiling," *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 379-384, 2012.

[3] D. V. Cicchetti: "Guidelines, criteria, and rules of thumb for evaluating normal and standardized assessment instruments in psychology," *Psychological Assessment*, Vol. 6, pp. 284-290, 1994.

[4] J. Cohen: "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37-46, 1960.

[5] G. W. Corder and D. I. Foreman: *Nonparametric Statistics for Non-Statisticians.* Hoboken: John Wiley & Sons, 2009.

[6] J. S. Downie, D. Byrd, and T. Crawford: "Ten years of ISMIR: reflections on challenges and opportunities," *Proceedings of the International Conference on Music Information Retrieval*, pp. 13-18, 2009.

[7] J. S. Downie, X. Hu, J. H. Lee, K. Choi, S. J. Cunningham, Y. Hao, and D. Bainbridge: "Ten years of MIREX (Music Information Retrieval Evaluation eXchange): reflections, challenges and opportunities," *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 657-662, 2014.

[8] J. L. Fleiss: "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, Vol. 76, No. 5, pp. 378-382, 1971.

[9] K. Hoashi, S. Hamawaki, H. Ishizaki, Y. Takishima, and J. Katto: "Usability evaluation of visualization interfaces for content-based music retrieval systems," *Proceedings of the International Society for Music Information Retrieval Conference,* pp. 207-212, 2009.

[10] X., Hu and N. Kando: "Evaluation of music search in casual-leisure situations," *Proceedings of the 5th Information Interaction in Context Symposium on-IIiX'14*, pp. 1-4, 2014.

[11] X. Hu, J. H. Lee, D. Bainbridge, K. Choi, P. Organisciak, and J. S. Downie: "The MIREX Grand Challenge: a framework of holistic user experience evaluation in music information retrieval," *Journal of the Association for Information Science and Technology*, under review.

[12] X. Hu and J. Liu: "Evaluation of music information retrieval: towards a user-centered approach," *Proceedings of the 4th Workshop on Human-Computer Interaction and Information Retrieval (HCIR)*, 2010.

[13] K. H. Krippendorff: *Content analysis: an introduction to its methodology.* Thousand Oaks: Sage, 2013.

[14] J. H. Lee and S. J. Cunningham: "Toward an understanding of the history and impact of user studies in music information retrieval," *Journal of Intelligent Information Systems*, Vol. 41, No. 3, pp. 499-521, 2013.

[15] J. H. Lee and R. Price: "User experience with commercial music services: an empirical exploration," *Journal of the Association for Information Science and Technology*, DOI: 10.1002/asi.23433, 2015.

[16] J. Nielsen: "Heuristic evaluation," In J. Nielsen, and R. L. Mack (Eds.), *Usability Inspection Methods*. John Wiley & Sons, New York, NY, 1994.

[17] S. Pauws and B. Eggen: "PATS: Realization and user evaluation of an automatic playlist generator," *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 222-230, 2002

[18] S. Pauws and S. van de Wijdeven: "User evaluation of a new interactive playlist generation concept," *Proceedings of the International Society for Music Information Retrieval Conference,* pp. 638-643, 2005.

[19] M. Schedl, A. Flexer, and J. Urbano: "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, Vol. 41, No. 3, pp. 523-539, 2013.

[20] F. Vignoli and S. Pauws: "A music retrieval system based on user driven similarity and its evaluation," *Proceedings of the International Society for Music Information Retrieval Conference,* pp. 272-279, 2005.

[21] H. De Vries, M. N. Elliott, D. E. Kanouse, and S. S. Teleki: "Using pooled kappa to summarize interrater agreement across many items," *Field Methods*, Vol. 20, pp. 272-282, 2008.