USER EXPERIENCE WITH COMMERCIAL MUSIC SERVICES: AN EMPIRICAL EXPLORATION

Jin Ha Lee
Information School, University of Washington
Mary Gates Hall, Suite 370, Seattle WA 98195
jinhalee@uw.edu

Rachel Price
Information School, University of Washington
Mary Gates Hall, Suite 370, Seattle WA 98195
rachelpr@uw.edu

Abstract

The Music Information Retrieval (MIR) community has long understood the role of evaluation as a critical component for successful information retrieval systems. Over the past several years, it has also become evident that user-centered evaluation based on realistic tasks is essential for creating systems that are commercially marketable. Although user-oriented research has been increasing, the MIR field is still lacking in holistic, user-centered approaches to evaluating music services beyond measuring the performance of search or classification algorithms. In light of this need, we conducted a user study exploring how users evaluate their overall experience with existing popular commercial music services, asking about their interactions with the system as well as situational and personal characteristics. In this paper, we present a qualitative heuristic evaluation of commercial music services based on Jakob Nielsen's ten usability heuristics for user interface design, and also discuss eight additional criteria that may be used for the holistic evaluation of user experience in MIR systems. Finally, we recommend areas of future user research raised by trends and patterns that surfaced from this user study.

Introduction

Evaluation is a critical component for successful implementation of any kind of Information Retrieval (IR) system. Recognizing this importance, the IR community has been running TREC (Text REtrieval Conference) since the early 90s to support IR research by providing an infrastructure for large-scale evaluation of text retrieval methods (Harman, 1994). In the Music Information Retrieval (MIR) domain, similar efforts were made through MIREX (Music Information Retrieval Evaluation eXchange), an annual evaluation campaign led by IMIRSEL (International Music Information Retrieval Systems Evaluation Laboratory) at University of Illinois. Since its conception in 2005, numerous MIR evaluation tasks have been carried out, and more than 963 individuals participated worldwide between 2005 and 2013[i]. This has significantly contributed to improving the performance of the algorithms/systems for specific tasks (e.g., music similarity, mood classification, artist identification, query by singing/humming) by allowing researchers to test their work on a large-scale music collection.

However, over the past 10 years, all the evaluation tasks have been focused on the "back-end" of the system, in other words, the algorithms and system performance. There is not a single evaluation task that is aimed to measure the "front-end" of the system such as the user interface, or more broadly, the user experience of the system. This is not surprising as the MIR field itself is relatively a young field, and therefore it was in fact critical to focus on the back-end, to test the applicability of existing algorithms to process music data, and explore ways to improve their performance. As the field matures, researchers are starting to recognize the "glass ceiling" effect on some tasks and subsequently are becoming more interested in how to improve the quality of user experience by working on the design of and user interactions with the system (Pachet & Aucouturier, 2004). The importance of considering aspects beyond the music content, such as the context of music and user is also being recognized (Schedl, 2013).

Moving forward in MIR evaluation, it is important to start exploring ways to evaluate systems considering a holistic user experience - the interface, the results, and also the ways users can interact with the music collection by listening, searching, browsing, and more. This is also evidenced by the announcement of the MIREX User Experience Grand Challenge 2014[ii] to celebrate the 10th running of MIREX, which is intended to evaluate the quality of users' experiences and interactions with web-based front-end music applications.

In response to the growing need for holistic evaluation of user experience in MIR, we conducted a qualitative user study exploring how users evaluate their experience with existing popular commercial music services, not only based on the results they get, but on their whole experience and interactions with music services on every aspect relevant to them. In particular, we aim to answer how users perceive their preferred music services meet baseline user experience heuristics, and which heuristics affect the user's decision to select a particular music service and continue to use it. Our findings may help improve the design of music services and also expand our understanding of user needs. In addition, we hope to contribute towards establishing a framework for evaluating the user experience specifically for music services.

Relevant Work

User-centered Evaluation in MIR

MIR studies on evaluation generally tend to be system-centric with a focus on the back-end, and evaluate how different algorithms or systems perform based on various attributes such as genre, artist, similarity, mood, etc. (e.g., Barrington, Oda, & Lanckriet, 2009; Lee, 2010; Urbano, Downie, McFee, & Schedl, 2012). With the recognition of the lack of user-centered research in the field, an increasing number of studies on user-centered evaluation were conducted (Lee & Cunningham, 2013). However, even in MIR studies on user-centered evaluation, users are typically asked to complete specific tasks providing feedback on perceived relevancy of songs, music similarity, mood, and so on. The main focus still tends to be on evaluating the performance of the system (i.e., how good they think the recommendations/playlists are) rather than their whole experience with the system. Our study instead aims to focus more on the front-end, but also looking beyond the interface and taking a more holistic approach by evaluating a wider variety of factors affecting how people experience music services.

Prior MIR studies that evaluate or seek users' opinions about existing commercial music services offer particularly relevant insights for our study. Barrington et al. (2009) evaluated Apple iTunes' "Genius" feature to two canonical music recommender systems, one based on artist similarity and the other on acoustic similarity. They discuss factors that influence playlist evaluation, including familiarity, popularity, perceived expertise of the system, and transparency, some of which will also be used as evaluation criteria in our study. Lamere (2011), in his blog article, also compares Google's Instant Mix, iTunes Genius, and the Echo Nest playlist engine, but mostly focuses on judging the quality of the playlists by identifying the irrelevant songs in them. The discussion hints on how personal preference and the context of music (such as which era the music is from) can affect the quality of user experience. Lee and Waterman (2012) also explore user requirements more generally for music information services. They have identified a list of 28 qualities including serendipitous discovery, inexpensiveness, convenience, customizability, etc., that users highly value. They do not discuss in detail, however, how each quality may or may not be used as an evaluation criterion for user experience. Our study aims to provide rich data about user context and use scenarios to complement the findings obtained through a large-scale survey like such.

Evaluating Usability and User Experience

There is a variety of methods for testing and evaluating usability and user experience within the field of Human-Computer Interaction (HCI). One common method is to judge the usability of an interface against a set of ten heuristics for usability, commonly referred to as Nielsen's heuristics (Nielsen, 2005). These heuristics are a baseline set of recommendations that are widely considered in the HCI field to be standards for "good" user experience (Nielsen, 2005). The heuristics have been used to evaluate the user experience in multiple domains such as e-government sites (Garcia, Maciel, & Pinto, 2005), ambient display design (Mankoff, Dey, Hsieh, Kientz, Lederer, & Ames, 2003), and medical devices (Zhang, Johnson, Patel, Paige, & Kubose, 2003). We included all ten of Nielsen's heuristics (2005) in our research as a baseline for the evaluation criteria for commercial music services.

- **Feedback**: Visibility of system status
- **Metaphor**: Match between system and the real world
- **Navigation**: User control and freedom
- **Consistency**: Consistency and standards
- **Prevention**: Error prevention
- **Memory**: Recognition rather than recall
- **Efficiency**: Flexibility and efficiency of use
- **Design**: Aesthetic and minimalist design
- **Recovery**: Help users recognize, diagnose, and recover from errors
- **Help**: Help and documentation

Another example of evaluation criteria for user experience is the model called

ResQue, consisting of 13 categories of questions designed to evaluate the quality of user experience on recommender systems (Pu, Chen, & Hu, 2011). This set of criteria is designed specifically with recommender systems in mind, and designed as a method for identifying essential qualities of a successful recommender system and factors that will motivate users to use the system. The framework has been adopted in a number of studies on recommender evaluation such as Cremonesi et al. (2011), Zhang et al. (2003), etc. Parts of the model that were deemed most applicable to our study include: quality of recommended items (accuracy), interaction adequacy, interface adequacy, perceived ease of use (ease of initial learning, ease of preference indication), and control/transparency. In addition, Herlocker, Konstan, Terveen, and Riedl (2004) also offer useful evaluation criteria beyond measuring the accuracy including novelty/serendipity and recommendation confidence. Knijnenburg et al. (2012) provides a larger framework for user-centered evaluation of recommender systems not only considering the user experience and interaction with the system, but also situational characteristics (e.g., privacy concerns, system-related trust, familiarity) and personal characteristics (e.g., domain knowledge, control). Tsakonas and Papatheodorou (2006) also investigate the characteristics of usefulness as a way of conducting a user-centered evaluation and suggest five usefulness criteria as important factors for user interaction: relevance, format, reliability, level, and timeliness. Some of these factors alight with criteria utilized in our own analysis, such as context (timeliness), confidence/trust (reliability), and relevance.

Although not all the music services/applications we examined are strictly classified as recommender systems, getting music recommendations was certainly a routine task carried out by many participants, and thus we decided to adopt some of the relevant criteria from the evaluation frameworks for recommender systems in our study, as shown below.

- **Recommendation Accuracy**: Items recommended match user interests (Pu et al., 2011).
- **Explanation**: Why items are recommended to the user (Knijnenburg et al., 2012; Pu et al., 2011).
- **Interaction Adequacy**: Recommender allows user to tell what they like/dislike (Pu et al., 2011).
- **Perceived Ease of Use/Familiarity**: User's acquaintance with or knowledge of the system (Knijnenburg et al., 2012; Pu et al., 2011).
- **Confidence/Trust**: User's belief that the system "works" (Knijnenburg et al., 2012; Pu et al., 2011).
- **Novelty/Serendipity**: Providing "non-obvious" recommendations (Herlocker et al., 2004).
- **Privacy**: Concerns about the amount and depth of information the user provides (Knijnenburg et al., 2012).
- **Overall Satisfaction**: How well the system fulfill the users' needs overall (Pu et al., 2011).

Research Design and Methods

For this study, a combination of two research methods was used: semi-structured interviews and "think-aloud" protocol. We wanted to adopt both explicit (ask) and implicit (observe) approaches (Herlocker et al., 2004). We provided a chance to participants to express why they choose to use particular services and how they evaluate the quality of the services in their own words. In addition, we also wanted to have an opportunity to objectively observe how they use particular music services and also learn more about their thought process as they interact with their preferred music services. The think-aloud method is commonly used to gather user data in usability testing (e.g., Nielsen, Clemmensen, & Yssing, 2002; Nørgaard & Hornbæk, 2006; Virzi, Sorce, & Herbert, 1993).

Participants recruited were all over 18 years old, who actively use at least one commercial music service/application which provides access to music content. Requirements for participation were that users had to have at least one preferred music service with which they were familiar. We did not limit recruitment of participants based on their stated preference towards any particular music system. All participants were undergraduate or graduate students at the University of Washington Information School. All the interviews were conducted between January and March 2014, either in-person or via Adobe Connect video conferencing. Each interview lasted approximately 45 minutes to an hour. A total of 40 participants were interviewed for this study. Participants were compensated with a $15 Amazon gift card for partaking in the study.

We asked the subjects what their most preferred music service is, with a focus on the music recommendation aspects of their preferred service. Eighteen participants chose Spotify, ten Pandora, three YouTube, two SoundCloud, and two Songza. Rdio, Bandcamp, Last.fm, Grooveshark, and iTunes were each chosen by a single participant. Subjects were interviewed about their preferred music services, specifically answering questions regarding the participant's interactions with the music services surrounding the system's music recommendation aspects, how they navigate the system and use the interface, the reasons why they prefer the music service(s) over others, satisfaction and frustrations they experience with the services, and how they interact with the services in a typical session. Subjects who use multiple services also shared their thoughts and opinions on other music services beyond their most preferred service. We wanted to gather data from a variety of users evaluating different types of services and see if we are able to apply a common set of evaluation criteria for their experience and interactions. Therefore, as previously stated, we did not limit our recruitment to users of specific music services. For future studies focusing on gathering larger amount of quantitative user evaluation data, it would make sense to recruit similar numbers of users per each service.

Participants were also asked to do think-aloud sessions, where they narrate their actions out loud to an investigator as they use their preferred music service as they would in a typical session. These tasks include known-item search, browsing albums, artists, or genres, interacting with recommendations, adding or deleting playlists, tweaking radio stations, etc. Both the audio and video of interviews and think-aloud sessions were recorded in order to document and be able to observe the interactions again as necessary.

Interviews were transcribed and coded by the contributing researcher. The codebook included categories based on Nielsen's ten heuristic evaluation criteria and other relevant criteria outlined above (i.e., recommendation accuracy, explanation, interaction adequacy, perceived ease of use/familiarity, confidence/trust, novelty/serendipity, privacy, and overall satisfaction). Additionally, codes representing seven different personas—archetypes representing the needs, behaviors, and goals of a particular group of users

(Cooper, 1999) epitomizing the most common types of people using music services—were also developed with a grounded theory approach (Corbin & Strauss, 2008). The discussion on the development of these personas will be reported in another manuscript currently under preparation. In order to maintain the consistency in the code application, we adopted a consensus model (Saldaña, 2009; Harry et al., 2005) where all the coded results by the first coder were reviewed the second coder, and any discrepancies or uncertainty in the code application were discussed until a group consensus was reached.

Limitations

This study is exploratory and qualitative in nature, and thus the objective is not to obtain findings that are generalizable to a larger population. Our goal is not to compare and rank systems based on the user data. Rather, we seek to obtain deeper insights into how users evaluate their experience. For instance, instead of simply identifying "privacy" as one of the evaluation criteria, we want to know which aspects of privacy users are concerned about as well as insights into their attitudes toward this criterion. We want to know which features of real commercial music services make them concerned about privacy issues. We envision that this will contribute toward building a user experience evaluation framework that may be used for future evaluation studies involving a larger population. We ask the readers to be cautious about interpreting the user feedback discussed here as representing the general public's opinions.

In addition, we note that the participants in this study were all students in Information Studies. On the surface it would appear that the results may be biased towards "savvy users" given the knowledge and perceived expertise of users in this study. However, many of the patterns that appeared from interviews in fact indicated that satisfaction with usability heuristics correlates more with persona types than with expertise of the system. These results are discussed in detail in another manuscript that is currently under preparation.

Findings and Discussion

Heuristic Evaluation for User Experience

Participants were asked a series of questions designed to understand how well they felt their preferred music recommender systems met basic guidelines for quality user experience. These questions were crafted based on Jakob Nielsen's ten usability heuristics for user interface design. Figure 1 presents part of the evaluation data on the top three most commonly used music services, illustrating the total user responses for each heuristic. Each icon represents a positive, neutral, or negative user response on a particular heuristic. Overall, all three of them seem to perform better on certain heuristics such as metaphor, consistency, memory, and aesthetics. Efficiency was the lowest rated heuristic across all three services, followed by prevention and navigation. Feedback, recovery, and help received many neutral comments. The consistent and predictable behavior of the system seemed to trump its efficiency or flexibility. This resonates with the observation that most participants were using each of these systems for very specific tasks, and often showed the "berry-picking" behavior of selecting the most appropriate system to serve different kinds of needs (Bates, 1989):

*"I use Grooveshark most of the time. I'll switch to Pandora if I'm looking for something new, but that only works for so long because the stations will start playing the same stuff over and over again. So, **my music life cycle is to hear a song, listen to it in Grooveshark over and over again, want to hear more stuff like it, create a station in Pandora for it, listen to that station, catch songs that I like on that, take them back to Grooveshark...Eventually I throw up my hands and buy it on iTunes.**" (User 23)*

*"If **Songza and Pandora had a baby together**, then I would use it." (User 33)*

There was also a strong satisficing (i.e., "good enough") theme emerging from the data, especially exhibited by participants using Pandora:

*"**As soon as I figured out the basics...**as soon as I found that I could look at some friends' playlists, and that I could find a few artists and make a radio station, I just, **I was like, I'm done. I'm done learning how to make this work.**" (User 1)*

*"There's nothing I don't like about Pandora...It might just be because I'm content enough...And I think I'm old enough, you know, I'm 45, **I'm not into that 'music is my world' type of mentality. So it's not high on my list.** Getting the oil changed in the car and making sure the laundry's done, the dishes done, the mortgage is paid, are a little bit higher priority than which music streaming application I'm going to use." (User 17)*

Spotify users, overall, seemed to desire more control and showed more willingness to spend time to curate their collection or provide input to the system.

*"[I don't use recommendations] at all...**I'm very self-directed in listening to music...I know what I want to listen to**, why am I just going to let a random radio station tell me? I actually forget that there's radio on Spotify." (User 8)*

*"**I feel like, strictly with Spotify, I want complete control.** That's my 'complete control' music interface, and Pandora is where I'm willing to let a little of that control go. So maybe that's why I am frustrated by the silly recommendations that they [Spotify] give me, because I'm like, **'No! This is the one place where I choose all the music, and you're not letting me or you're trying to give me different suggestions.'"** (User 12)*

This trait was also exhibited in users who used less mainstream music services such as SoundCloud or Bandcamp:

*"For me it's not really worth the time. I think it's just going to recommend stuff that's also tagged [similarly]...**I do my own ways of [finding], and I rely on my friends and people I write with to recommend stuff**...It's just **a fun thing to find new stuff**..."(User 6 on SoundCloud)*

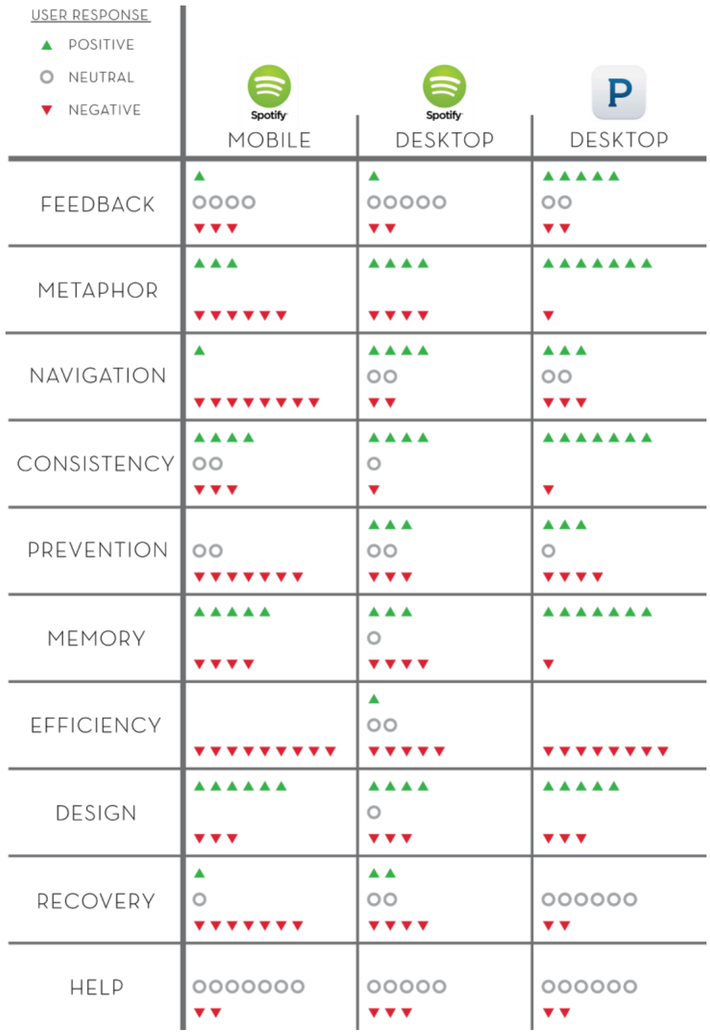

Figure 1. Comparison of ten heuristics for the three most commonly used music services: Spotify Mobile, Spotify Desktop, and Pandora Desktop[iii]

Feedback

For this heuristic, we sought to understand if users knew what the music service/application was doing at any given time; was it playing a song, and if so, which one? What is coming up next? If the system is not playing a song, why not? Questions posed to participants included: "How well do you feel the system informs you of its status? Do you feel like you always know what is going on, when the system is working normally or when it is not working normally?"

The interruption during a task or the lack of the visibility of system status often led participants to feel negatively toward particular music services.

> *"I guess I usually know what's going on. The ads do pop up unexpectedly, but I guess that's how ads always are.* ***Especially if you're in the middle of doing something, you're trying to interact with it, and then an ad pops up, that's confusing.****" (User 5 commenting on Pandora)*

*"I remember going to the website and **trying to read all these FAQ's and help, and it was just not working**...What happened was that I was doing it right, but it was just taking time [but it didn't tell me that]...**I think all I really had to do was log out and log back in, but I didn't know that,** because the system didn't tell me." (User 27 commenting on connecting Spotify to their Facebook account)*

Overall users responded neutrally or positively about this heuristic; their preferred music systems generally did a "good enough" job of indicating system status. Pandora (web interface) had more positive responses than other systems, due to the "always on" centrally located play status bar located at the top of the screen. Spotify's mobile app received more negative responses than other systems; users commented that it was not always easy to tell what was playing, or if the system was not playing anything, why the system was not working. Some participants also pointed out how the experience varies depending on the platform:

*"I feel like it does a really good job of showing me what's going on on the website, but **not always on the mobile**. Like, I can see what was previously playing on the web, but I can't see that on mobile. I do like how right now, how it shows me that it's paused [points to big "paused" button]. I like this whole set up [points to static play panel]." (User 29 commenting on Pandora)*

Metaphor

For our context, we wanted to know if users felt that the labels used in navigation menus were clear and easy to understand, if text-based content on the site was written in a meaningful way, and if navigational concepts were easy to understand. Questions posed to participants included: "Does the language that the system uses to label and describe things work with your own natural language? How so, or why not? Do you feel like they are speaking your language, or are these words written for someone else? If you see error messages or labels or other content, do you understand the message? Do you ever find the language vague, misleading, or confusing?"

Overall the participants seemed to adequately understand the meaning and intentions of labels, with a few recurring exceptions. Pandora users felt that the labels and language used in Pandora's web interface were easy to understand. There was little confusion about the function of menu items based on their labels. Spotify fared a bit worse; users who gave negative feedback for Spotify (mobile or desktop) indicated that some of the navigation labels on Spotify's primary navigation was vague – a common comment was that the difference between "browse," "search," and "discover" as navigation labels was unclear.

Participants who strongly expressed frustration were those not interested in investing a lot of time and effort to fully understand the all the features of the system.

*"Some of the [navigation] tabs are kind of vague. On the navigation, **there is 'search' and 'browse', which can mean the same thing**. There's been some times when I'm like, "I don't even know what that means"...Like, I'm trying to save a song...and **you can***

"star" an artist or "save" them on a playlist. I don't know what the difference is*...There's a lot of overlapping. I tend to just stick with what I know." (User 26 on Spotify).*

Navigation

We sought to understand if users felt like they were able to move forward and backward as they navigated through the service/application, if they could undo actions that they had taken by accident, and if they were able to retrace their steps to adjust their search or browsing path when they made a mistake or changed their mind about the path they had chosen. Questions posed to participants included: "What do you do when you make a mistake while using the service? Is it easy to back up to your last step? Do you have to start over? Have you seen an 'undo' or 'redo' option in this interface, and if so, do you use it?"

Generally, the mobile users felt less control than desktop users. Spotify (desktop) and Pandora (desktop) received mostly positive or neutral feedback. Users felt like they knew how to undo actions they had not meant to complete, or responded neutrally, indicating that they had never run into a problem that they could remember. Spotify (mobile) fared quite poorly on this heuristic; users commented that they would often accidentally hit things with their thumbs due to the limited screen size, but would not be able to undo their actions if necessary.

> *"**I never know how to undo [a change], and so I'm so loath to touch anything on this screen, because it will screw with everything.** I wish I had more control over that. I wish it was 'I organize it how I want it'. I feel like it's curating my music and I'm accessing it, and that's annoying." (User 19 on iTunes)*
>
> *"Yeah, so the thumbs-upping and thumbs-downing thing is, **they're so close together, so it's easy to slip and hit the wrong one**. I don't recall getting any sort of ability to undo, or if there was, it wasn't as straightforward as I would have liked...I've definitely thumbs-upped things...well, I think you can undo that. But if you thumbs-down then it stops playing and moves on to the next song, and I think you can go back, but it's not as obvious, so that one is like, **'Oh, I actually liked that song, I don't want it to not play'.**" (User 22 on Pandora)*

Consistency

In our context, we focused on consistency in interactions. Users were asked about the "predictability" of interactions and menu placements across the interface. We also sought to understand how well systems followed "platform conventions", i.e., how well they used metaphors already commonly understood from other music services/applications in their own design. Questions posed to participants included: "Do you feel that the labels, words, descriptions, and actions in the system are consistent across tasks? For example, is it the same process to build a playlist or channel as it is to browse albums? Do you feel that if you understand how to manipulate one part of the system, you are confident manipulating other parts of the system? For example, if you know where a menu will 'hide' in one part of the service/application, do you know where to find similar menus in other parts? Is there any part of the system that hangs you up because it behaves differently than you expect? Is it consistent with itself, and is it consistent with other services you have seen?"

Overall, users responded positively, and felt that their preferred systems maintained consistency across menu functions and other interactions. Additionally, users felt that their systems were built in ways that were approachable and modeled after other well-known music players, like iTunes or YouTube. They felt that by understanding how to use a particular part of the system, they could figure out how to use most other parts. Spotify mobile users indicated that some menus were hard to find due to interactions required to activate them, and some screens were formatted differently than others. .

> *"It's usually decently consistent. Like, even when they do change up how to play a song, it's in a way that's already kind of set out, **like YouTube**. Or on the discover page, it's a big triangle that you hit [to play] instead of [clicking on] a song, so it incorporate album art there, instead of having the iTunes [model] kind of list and metadata. So, that is different, but it's not necessarily bad. You can add songs [to playlists] **like you do in iTunes**, by dragging them or by right-clicking." (User 16 commenting on Spotify)*

> *"One thing that is annoying to me and that **I've never been able to figure out how to fix...[clicks around in several menus]...is to change the display.** Some places it shows up as gridded, and some places it's a list...and I thought you fixed it here...I guess not." (User 19 on iTunes)*

Prevention

This heuristic translated directly to our context, where we sought to understand how users felt systems allowed them to avoid making mistakes, such as deleting playlists or radio stations, accidentally creating new playlists or stations, or making other changes. We also wanted to see if systems had a habit of requiring confirmation to commit actions that might have been committed by mistake. Questions posed to participants included: "Does the system 'warn' you when you're about to make a mistake? What types of mistakes have you been warned against? (e.g.,: deleting a channel/playlist/song) Can you think of a time when you made a mistake and wishes that you had been paused for confirmation before clicking 'ok'?"

All the systems we examined received a majority of negative responses regarding error prevention. While it seemed more difficult for users to vocalize their opinions around this heuristic, many users responded negatively. Users indicated that they did not feel the system guided them in such a way as to prevent them from making errors, and that if the system provided more feedback or confirmation messages for actions, that they may be less likely to make common errors. Some users suspected that the systems were trying to "stay out of the way", which they admired, but which sometimes resulted in errors.

> *"I think the interface doesn't want to get in the way. it would be nice if the interface would expose some of its logic, for me anyway. I would like to know, "You clicked the skip button", but it doesn't [confirm my feedback]. **If I like a song, what happens? If I dislike a song I meant to like, I just don't know what it's doing."** (User 5 on Pandora)*

*"There is an undo option, but I've never used it. **I've never made a detrimental mistake, so that seems pretty good for their system.** [user tries to delete playlist]. It asks me if I'm sure if I want to delete the playlist." (User 16 on Spotify)*

Memory

In this context we sought to understand how much of users' actions or paths were memorized or fostered by the design of the interface. We also wanted to understand how easy it was to accomplish primary (routine) tasks and secondary tasks. Questions posed to participants included: "How visible are all of your options and menus when you use this system? Do you find yourself trying to remember how to do a certain action, or is the interface built in such a way to help you 'find the right button'? Are your primary tasks (those that you do daily) easy to accomplish with very few clicks or menus?"

Users were split between positive and negative responses, with the exception of Pandora desktop users, who overwhelmingly responded positively. Users indicated that the majority of their primary tasks (playing music, accessing saved playlists) were surface-level, and were clearly presented within the interface. They did not have to "dig" through layers of menus to accomplish most tasks. Secondary tasks, those that are not as frequently completed, such as deleting playlists, following "related artists" links, or adjusting account settings, may take additional steps to accomplish but are still relatively close to the surface. Some respondents indicated that there are features of their preferred music systems that they have never used or explored because they are not readily presented at the surface-level. For example, many Pandora users had never looked at their "account" option or "profile", because they did not know it even existed until they actively explored all of the menus during the interview.

*"Every once in a while I have to kind of figure out, or **remember how to do something that I thought I knew how to do.** [Laughs]. I'd say I've memorized the process for most things. Or if I haven't, then I just know that I'm in for a little bit of clicking around and trying to find the right menu or something. [Points to screen] Like, the fact that I can't figure out how to fix the view of this list of albums is a little frustrating [clicks several menus trying to figure it out]." (User 19 on iTunes)*

*"I just remember what to push. So…it's using both words and images [in the menus]. It's hard to say how helpful those images are. [Looks at menu]. There's not a lot of options...**There's nothing in these icons that really indicates what it's going to do...I know because I have clicked and I know what it will do**." (User 8 on Spotify)*

Efficiency

This heuristic required a bit of translation for our purposes. We framed this in the context of customizing home screens, playlist displays, and primary menus for accelerated access to primary tasks, as well as utilizing keyboard shortcuts for common tasks (play, pause, skip, thumbs up, thumbs down). Since each user comes to a service with a distinct set of routine tasks, each user has a unique "wish list" of menu items that, in a perfect world, would be always accessible and top-of-mind. Customization, in this case, might consist of only showing certain menu elements, reordering menu options, and deciding which stations or playlists are always displayed for a user to choose from quickly. Questions posed to participants included: "Depending on how frequently you use the system, would you call yourself an expert or a novice on this system? Do you use it differently now, after much experience, than you did when you started? Do you use keyboard shortcuts, have you personalized the homepage, etc., that makes it easier for you to do the things you need now than when you first started using it?"

All systems received overwhelmingly negative responses for flexibility. Users felt that they could not adequately customize their experiences to suit their needs. This was in fact the worst-performing heuristic of the ten. Users wished that they could organize playlists or stations based on their own criteria, by nesting them hierarchically or tagging them, or organizing by "most listened to" or "most frequently used". Another example of customization dealt with the primary navigation. Many users approach their systems with one primary need, but systems such as Spotify are built to accommodate a wide variety of needs. Users wished that they could customize their static navigation menus to suit their specific needs, rather than having to scroll through all options to get to the one option that they use most. For example, many of the Spotify users rarely utilize the main navigation components that are kept static on the left hand side of the screen, and have to scroll down to get to their playlists. Users expressed a desire to drag and drop this left hand navigation to suit their purposes; for example, dragging the "Playlist" portion of the navigation to the top of the page, since that is the action they complete most frequently and want to access quickly upon opening the interface. This type of customization was deemed important by participants, but most systems do not currently accommodate this need.

> *"**I would love to hide [options I never use].** These sorts of things I wouldn't want on my main menu, I might want them under 'options'. There's quite a bit that I would hide. Like, my desktop, I normally only have what I really need. So **I would make it more customizable** as far as making it more minimalistic. If they had a 'lite' version, I would use that more...like the iTunes mini-player, for example." (User 13 on Spotify)*

Design

This heuristic translated easily. We sought to understand how users felt about the aesthetic aspect of their preferred service, and sought to understand specifically if users felt the designs were clean or minimalistic. Questions posed to participants included: "How do you feel about the design of the interface? How would you describe it, in your own words? What does it look like, and how does it feel (e.g., crowded, clean, cluttered, bright, dark)?"

While some users felt that the interface of Spotify was cluttered, most users felt that the interfaces of their systems were well-proportioned and clean. A common complaint was that advertising space marred the aesthetic effect of the interface. However, most users also understood that they got what they were paying for; those who were using free accounts had learned to tolerate ads that made the interface cluttered and distracting.

> *"I like [the aesthetics]. I feel like **they're pretty unobtrusive, and it's easy to find the***

> ***main things that I want****, which are my radio stations, what is it playing, and where is the player. I feel like it works pretty well...I would get rid of the ads." (User 29 on Pandora)*
>
> *"[The aesthetic] is photo-heavy on the homepage...and then once you get into the content, it's content-heavy. So either way you are getting a lot of either visual or audio content. It's not busy...They understand that music is a force that fuels other things you do, and so also with the language they use, and with these images,* ***they don't make assumptions about who their audience is. The design of the interface stays out of the way.****" (User 32 on SoundCloud)*

Recovery

This heuristic applies directly to music systems. We sought to understand if error messages were easy to understand, written clearly, and suggested solutions for fixing the problem. Questions posed to participants included: "Think of a time or a situation using this system when you received an error message in response to your action. When you do make a mistake and get an error, how do the error messages read? Are they written in a language that helps you understand what the error was? After reading an error message, do you know how to resolve the problem or what to do next?"

The general response to this heuristic was that error messages tend to be uninformative, vague, and do not often present a solution to the problem. Some users could not think of a time that they had seen error messages at all. Typical scenarios resulting in error messages were results of connection issues. In some instances, the program would stop working and would not provide any error message at all, in which case users typically abandoned the system.

> *"****It doesn't tell me what's wrong it just...doesn't work.*** *I don't know if the error messages are always understandable. I don't always get an error message when something goes wrong. I feel like there's only one error message [regardless of the problem]." (User 26 on Spotify)*
>
> *"Sometimes I get into this 'offline mode' and then it won't let me on, and I have to close the system down and restart my computer.* ***The error message doesn't do a good job of telling me how to fix it****...I don't think it ever tells me to restart, but I just know that from working with computers." (User 16 on Spotify)*

Help

This heuristic translated directly to understanding how users feel about the help documentation of their preferred service, if they had ever used it. Questions posed to participants included: "Think of a time when you got 'stuck' and couldn't figure out how to accomplish what you set out to do in the system. When you can't figure something out on your own, do you think there's adequate documentation/help to assist you in figuring out how to complete your task? Do you use help documentation on this system? Do you seek out other avenues of getting help? What happens if you can't find what you need right away?"

Most users had never tried to use the system's help documentation. When asked why, users responded that it was not necessarily because they never had a problem, but because they normally turned to external sources before trying internal documentation. Some would ask friends or relatives for help or would simply conduct a Google search. Some users even admitted that while they may have ended up using help documentation from the service itself, they were likely to have accessed it via Google search rather than from the interface.

> *"That was a pain.* ***I didn't even know how to phrase what I was looking for****, and so I would type in what I thought were keywords... I'm sure it's in there somewhere, I just couldn't find it because I wasn't using the right search phrase. By nature, I just try to type natural language, like, 'How do you do da-da-da,' and I wasn't really getting what I was trying to get at.* ***I think I actually found it by doing the same search in Google*** *and then I found [the answer] in a forum. So Spotify wasn't really helping." (User 27 on Spotify)*

> *"I have not been able to figure things out [in the system] before,* ***but I don't think I've ever used the 'help'. I just stopped trying to do it.*** *I think if it offered [help] when I was trying to do something, like if something popped up, I'd like that. For a system like this,* ***I'm not interested in investing that much time in figuring it out.*** *It's kind of like, I'm coming here because* ***I want something relaxing, something that's easy, or for music discovery, something that's fun or interesting,*** *and if it's not going to give that to me then I'll go somewhere else." (User 29 on Pandora)*

Other Evaluation Criteria

In addition to Nielsen's ten heuristics, we adopted eight other criteria for evaluating the quality of user experience for recommender systems in order to better understand what qualifies as positive user experience within music services.

Recommendation Accuracy

This criterion investigates how well users feel that recommended items match their interests. We asked users to provide their perceived "success rate" for "good" recommendations, and to explain their answer. No users claimed that their systems were 100% successful; many users responded that the recommendations were around 50% "accurate" or "successful" in recommending "good" items. Definitions of accuracy did change among users. Some indicated that it meant they recognized the artists but may not have listened to them before, other users indicated that it meant the recommendation fit in well with other songs they were currently listening to, or that it did not "stick out" enough to warrant negative feedback. Some users did not like the attributes/relationships that systems based their recommendations on:

"Listening to Aloe Blacc would not make me want to listen to that [recommendation], and so I just don't trust their judgment. ***I'm just like, why? Stop trying to tell me how to live my life. I don't think that's a good suggestion for me at all!****" (User 12 on Spotify)*

Other users thought that their recommendation system did a good job of understanding more complex relationships between artists beyond musical similarity, resulting in better recommendations:

"75% [accuracy rate for recommendations]. I think it's something they do very well. It's really tough, because sometimes, ***it's not about 'Oh their music sounds the same as these guys'...If I find a band I really like I don't want to hear the exact same music from a different band.*** *So it's kind of like finding the music that goes well with this [other band]." (User 24 on Spotify)*

Explanation

This criterion asks users if the system explains why items are recommended to them; if it provides adequate explanation of the reasoning behinds its recommendation. Most users indicated that they would like to see more explanation behind the logic of the system's recommendations than is currently available to them. In some functions of Spotify like the "Discover" function, Spotify gives vague explanations for recommendations, such as, "Because you listened to X Artist you should try listening to Y Artist." Users indicated this was a good start, but even more information would be better. Pandora users mentioned that Pandora does display some of the criteria used for its recommendations.

*"****Sometimes I wonder why things are on there. I guess I need more insight on why I should choose to click on this thing****...if it's a band I've never heard of I'm not going to click on it unless there's a reason for me to. So I need more about what they sound like. It should only be [based on] bands that I have listened to often. I feel like it should just [pull] from songs that I've listened to the most. A lot of times it's like,* ***'You listened to this song by Rihanna once. All of a sudden we think you should listen to Justin Bieber.' That doesn't work for me.****" (User 31 on Spotify)*

Interaction Adequacy

This criterion asks users if the system allows users to provide feedback by indicating what items they do or do not like. Further, we asked users if the depth of feedback they could currently provide was enough. For systems that do allow feedback, many users indicated that they wished they could provide more detailed feedback than a simple "thumbs up" or "thumbs down". Users responded that they would, for example, like to tell the system specifically why they thumbed an item down:

"One day I was listening to [a station] and I was making a ***concerted effort to not have it play me things that were too bluegrass-y,*** *so I disliked [thumbs-downed] a*

few tracks... And ***what it seemed to think is not that I did not want bluegrass, but that I did not want vocals.*** *So for the rest of the time I was using the station it played me nothing but instrumental bluegrass. It seems like it's making assumptions [about my feedback]." (User 5 on Pandora)*

Some systems do not allow feedback, like Spotify's "Discover" functionality. Users indicated that they would be willing to improve Discover's functionality with feedback, whether by whitelisting or blacklisting recommendations:

"I really like the artist Lykke Li, and they always recommend the artist Ellie Goulding, but I hate Ellie Goulding. ***Even though their genre is comparable,*** *like, female electropop…[I wish I could tell it]* ***never recommend this to me again.*** *No, Spotify, I don't want that, and stop recommending it!" (User 15 on Spotify)*

Perceived Ease of Use/Familiarity

This criterion seeks to understand how easy it was for users to learn their adopted system, and how familiar they were with the system at the time of our research. Most systems were easy to learn. Some users indicated that more complex systems, like iTunes or Spotify, were harder to master.

*"****Spotify's interface was just really different, and I didn't understand it, and I don't adapt to change very well…****I've had the app for several months, and just yesterday did I actually feel like I had a handle on creating a playlist seamlessly without getting lost. It's clunky and it's not intuitive." (User 3 on Spotify)*

Many users considered themselves "experts" at the time of the interview, most logging 2 or more hours per week on their system for 6 months or more and remarked upon the ease-of-use of their preferred system:

It's pretty easy to use overall*, and* ***it does what I want it to do, which is play music.*** *It's easy for me to accomplish the tasks that I need to do on a daily basis."(User 22 on Spotify)*

Confidence/Trust

This criterion asks users if they trust the system and the recommendations it makes. It is vaguer than some criteria and may overlap with privacy concerns. We sought to separate this from privacy by asking users about their belief in the system's ability to provide recommendations that would rival the recommendations users might get from other people, be they friends or strangers. Some users were confident that their preferred system could make accurate recommendations:

"I think it does a really good job. So, for example, I know that if I'm listening to "Today's Country Radio" station, I know I'm going to get a lot of popular top 40

country hits, and that's what I expect. I know that if I listen to Gary Allen radio, not only do I get Gary Allen…I also get ***some of the not-popular artists too, so they do a good job of blending those.*** *But I think they do a good job of having similar artists." (User 12 on Pandora)*

Other users were doubtful that a music recommender could ever make recommendations as "good" as those that would come from another person:

"I like seeing ***the feed [of what my friends are listening to]. That's usually more reliable than the algorithm****, I think I have a list of* ***friends that I really respect their listening habits and music tastes.*** *I usually trust that pretty well." (User 16 on Spotify)*

Novelty/Serendipity

This criterion investigates whether or not users feel that the system provides "non-obvious" recommendations for items the user may not otherwise have discovered. For our purposes we did not seek to understand the fine differences between novelty and serendipity as described in detail in Herlocker et al. (2004), but merely if users felt that the system did a good job of providing either novel or serendipitous recommendations. Users who expressed high levels of musical expertise or discerning taste tended to respond their recommender frequently fell short in providing novel recommendations:

"I don't like: [Spotify recommends:] 'Holiday – Boulevard of Broken Dreams' was popular while you were in high school, play now?' I could see how that could work with some people, but that wasn't popular with me, so that's not that useful…[points to several recommendations of bands] I already know all these…[My judgment] is ***dependent on whether I know the bands. If I don't know them at all, I'm probably more interested in clicking on it*** *because it's someone that I could discover music in that way instead of just saying 'Oh, these artists are related.' If it was [a band] I knew, I probably wouldn't click on [the recommendation] because I already know that album." (User 16 on Spotify)*

Users who tended to be more "adventurous" or who admitted to spending little to no time actively curating their own musical experiences responded positively:

"There are some times when it recommends me things that I never would have thought of, and so I think, 'yeah, I'll give it a shot.' And I listen to it, and I think, 'Oh, this is pretty cool!' ***For me, I'm generally pretty adventurous with my listening habits,*** *so I'm always like, 'yeah, sure, I'll check this out'." (User 11 on Last.fm)*

Privacy

This criterion seeks to understand the privacy concerns of users, and how those concerns affect the amount of information users will provide to improve recommendations. We asked users about their privacy concerns as they relate to the ability of a system to provide better recommendations based on more personal information. Users provided a variety of answers, with varying degrees of openness or concern about the security of their personal information. Some users expressed strong distrust or conflicted feelings in regard to privacy issues:

> *"I already feel like a lot of places have too much information about me...* ***I wouldn't want to give a system more information about me even if it would provide a perfect playlist,*** *because I still want to have control of that [information]…* ***It's creepy.*** *I like having some degree of control and privacy." (User 13 on Spotify)*

> *"I'm split between 'that's really cool' and 'that's kind of creepy'. If I had the option to control [the information it has about me] then that might be something I accept. I still want the option to work outside of what they can gauge about me,* ***I don't want the system to make assumptions.*** *But* ***at the same time I think that it's really neat*** *that that is a possibility." (User 30 on Pandora)*

Overall Satisfaction

This criterion asks if the user is satisfied with the system overall. In a sense, this is the ultimate higher-level criterion which determines if the user is going to continue to use a particular system or move on to using something else. Users were asked to rank their satisfaction on a scale from 0-10, zero being "extremely unsatisfied" and ten being "extremely satisfied". Answers were scattered and ranged from 4-9.5 (e.g., mean of Spotify = 7.29, Pandora = 7.2, and YouTube = 8.3).

> *"[Ranked the system at 9.] The user interface is good. It's to-the-point, I can find what I need, I can see the options that are available…* ***I cannot conceive of anything else that I would want in there."*** *(User 17 on Pandora)*

> *"It's a step below neutral. So, 4. It's annoying enough that* ***I have a desire to leave, but it's difficult to leave****…The amount of cognitive overhead for me to try to [switch services], when all I really want to do is spend a few minutes finding something to just listen to and just keep in the background…[makes switching] difficult." (User 35 on Spotify)*

We do not consider these numbers as representations of how the general public feel about these services due to the small number of participants. This will be explored in our future work where we survey a larger user population.

Conclusion and Future Work

In this paper, we used a framework built upon known heuristics such as Nielsen's and expanded by evaluation criteria around privacy, transparency, and trust, to evaluate the user experience of commercial music services. A number of themes surfaced from our research.

Satisficing/"Good Enough"

Many users mentioned that the service was "good enough" for their purposes. Maintaining the consistency, in other words, conforming to users' expectation on how they think the system should behave based on their previous experience, was critical. Many expressed that a "change" would be difficult. Once they felt "committed," users were generally quite forgiving of some of the negative aspects of the services. This aligns with findings in other user evaluations of interfaces that "usefulness precedes usability" (Tsakonas & Papatheodorou, 2006). When we asked about what would make them consider switching to another service, most participants replied the new service would have to seamlessly "translate" all of their saved playlists and favorites to the new service to make up for the time invested in the original service.

Berrypicking Search Behaviors

Many participants exhibited "berrypicking" search behaviors (Bates, 1989), using a variety of services to accommodate their needs as they arise, such as supplementing Pandora listening with YouTube searches. Their searches often changed and evolved, and users were able to select the best service to user for different stages in their searches. This suggests that services supporting very specific use cases may be more reasonable than attempting to develop a "one-size fits all" type of service that supports a number of use scenarios. As previously mentioned, most users did not spend a lot of time and effort trying to even explore all the features provided in these services.

Different User Attitudes

The participants exhibited very different personalities and attitudes toward music services. Some clearly viewed these services as "tools" to help them, and explicitly said they do not expect the machine would truly understand their tastes or should "try to tell them how they should live their lives." Some were more open to relying on the recommendations provided by the systems for serendipitous discovery of music, or spent a lot of time curating their listening experience. On the other hand, there were also participants who clearly had no interests in providing any sort of input and expect the system to just tell them what they should listen to. Different types of music services seemed to appeal to different types of users (e.g., Pandora users want to spend minimal amounts of time interacting with the system vs. Spotify users, who are willing to spend more time curating their collection and shaping their music listening experience), and thus it will be critical for developers to understand the general "personalities" of their users to identify the best design strategy.

For our future work, we plan to tease out these different persona types (e.g., guided listeners, non-believers, music epicurean) that may provide additional insights on how to design music services for targeted user populations (Lee & Price, under preparation). In addition, we plan to conduct a survey asking a larger user population about their use of a wider variety of music services and how they would evaluate their experience based on the framework established in this study.

Acknowledgment

This research is supported by the University of Washington Information School Research Seed Fund.

References

Barrington, L., Oda, R., & Lanckriet, G. R. (2009). Smarter than Genius? Human Evaluation of Music Recommender Systems. In *Proceedings of ISMIR 2009*, 357-362.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Information Review*, *13*(5), 407-424.

Cooper, A. (1999). The inmates are running the asylum. Indianapolis: Sams.

Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Thousand Oaks: Sage Publications.

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. V., & Turrin, R. (2011). Looking for "good" recommendations: A comparative evaluation of recommender systems. In *Human-Computer Interaction–INTERACT 2011* (pp. 152-168). Springer Berlin Heidelberg.

Garcia, A. C. B., Maciel, C., & Pinto, F. B. (2005). A quality inspection method to evaluate e-government sites. In *Electronic government* (pp. 198-209). Springer Berlin Heidelberg.

Harman, D. (1994). Overview of the second text retrieval conference (TREC-2). In *Proceedings of the workshop on Human Language Technology* (351-357). Association for Computational Linguistics.

Harry, B., Sturges, K., & Klingner, J. (2005). Qualitative data analysis: Mapping the process. *Educational Researcher, 34*(2), 3-13.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5-53.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441-504.

Lamere, P. (2011). "How good is Google's Instant Mix?" http://musicmachinery.com/2011/05/14/how-good-is-googles-instant-mix/.

Lee, J. H. (2010). Crowdsourcing Music Similarity Judgments using Mechanical Turk. In *Proceedings of ISMIR* 2010, 183-188.

Lee, J. H., & Cunningham, S. J. (2013). Toward an understanding of the history and impact of user studies in music information retrieval. *Journal of Intelligent Information Systems*, *41*(3), 499-521.

Lee, J. H., & Price, R. (under preparation).

Lee, J. H., & Waterman, N. M. (2012). Understanding User Requirements for Music Information Services. In *Proceedings of ISMIR 2012*, 253-258.

Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., & Ames, M. (2003). Heuristic evaluation of ambient displays. In *Proceedings of the ACM SIGCHI conference on Human factors in computing systems,* 169-176.

Nielsen, J. (2005). “Ten Usability Heuristics.” http://intra.iam.hva.nl/content/1112/verdieping1/research_for_design/intro-en-materiaal/RfD-Heuristic-Evaluation.pdf.

Nielsen, J., Clemmensen, T., & Yssing, C. (2002). Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction,* 101-110.

Nørgaard, M., & Hornbæk, K. (2006). What do usability evaluators do in practice?: an explorative study of think-aloud testing. In *Proceedings of the 6th conference on Designing Interactive systems,* 209-218.

Pachet, F., & Aucouturier, J. J. (2004). Improving timbre similarity: How high is the sky?. *Journal of Negative Results in Speech and Audio Sciences*, *1*(1), 1-13.

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems,* 157-164.

Saldaña, J. (2009). The coding manual for qualitative researchers. London: Sage.

Schedl, M. (2013). Ameliorating Music Recommendation: Integrating Music Content, Music Context, and User Context for Improved Music Retrieval and Recommendation. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia,* 3.

Tsakonas, G., & Papatheodorou, C. (2006). Analysing and evaluating usefulness and usability in electronic information services. Journal of information science, 32(5), 400-419.

Urbano, J., Downie, J. S., McFee, B., & Schedl, M. (2012). How Significant is Statistically Significant? The case of Audio Music Similarity and Retrieval. In *Proceedings of ISMIR* 2012, 181-186.

Virzi, R. A., Sorce, J. F., & Herbert, L. B. (1993). A comparison of three usability evaluation methods: Heuristic, think-aloud, and performance testing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 37, No. 4, pp. 309-313). SAGE Publications.

Zhang, J., Johnson, T. R., Patel, V. L., Paige, D. L., & Kubose, T. (2003). Using usability heuristics to evaluate patient safety of medical devices. *Journal of biomedical informatics*, *36*(1), 23-30.

Zhang, Y. C., Séaghdha, D. Ó., Quercia, D., & Jambor, T. (2012). Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining,* 13-22.

---

[i] http://www.music-ir.org/mirex/wiki/MIREX_HOME
[ii] http://www.music-ir.org/mirex/wiki/2014:GC14UX
[iii] Other services omitted due to small number of users who selected them as the most preferred music service.