Play it by Ear: A Case for Serendipitous Discovery of Places with Musicons

Louis Lei Yu, (lyu@gustavus.edu)

Anupriya Ankolekar (anupriya.ankolekar@gmail.com),

Thomas Sandholm (thomas.sandholm@gmail.com),

(This work was finished when Anupriya Ankolekar and Thomas Sandholm were both researchers, and Louis Lei Yu was a postdoctoral research fellow at Hewlett Packard Labs.)

Human-Computer Interaction (HCI)

- Not my research area (but I like doing research projects with my friends, so why not?)
- The design and use of computer technology, and the interfaces between people (users) and computers.
  - Faster, more efficient, more natural and/or intuitive .

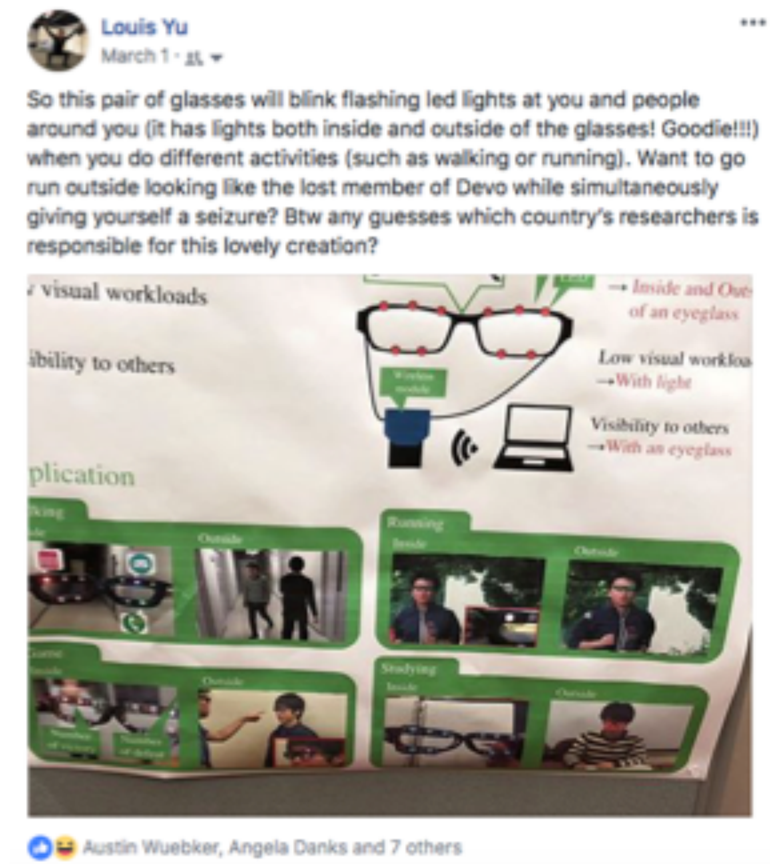
Play it by Ear, a Case for Serendipitous Discovery of Places with Musicons, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)

Evaluating Mobile Music Experiences: Radio On-the-Go, in Proceedings of the 9th EAI International Conference on Mobile Computing, Applications and Services (MOBICASE 2018)

Foxtrot: a soundtrack for where you are, in Proceedings of the Interacting with Sound Workshop: Exploring Context-Aware, Local and Social Audio Applications (IwS 2011)

Context Aware Signaling

- Signals in certain context that gives you extra information in non-intrusive ways.



Setting the Mood

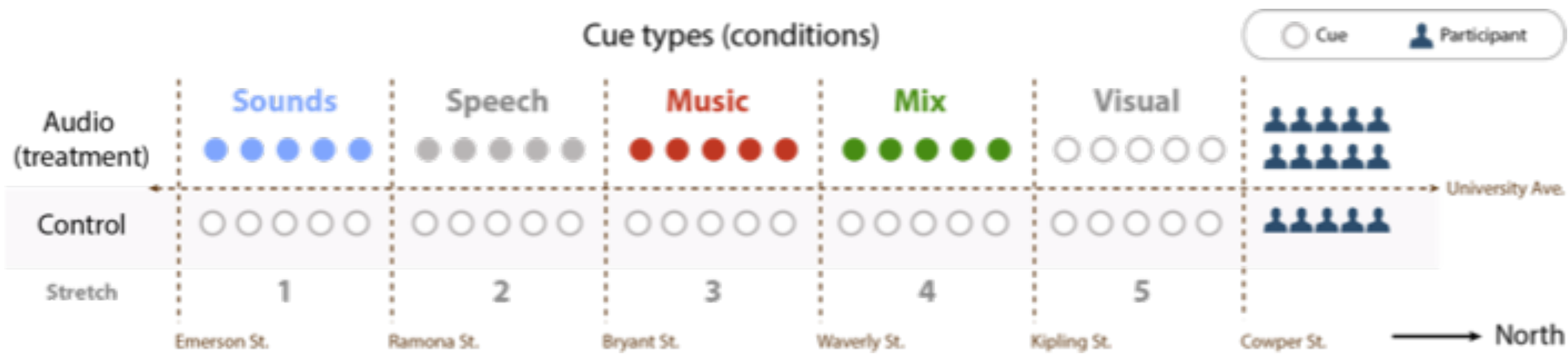
- Setting: you are lost, and walking on the busy street in an unfamiliar city.
- You have your iPhone/iPod with you.
- You are curious about nearby places.
- I will play you 4 sound samples. For each, close your eyes, listen for a few seconds, then write down the first location that comes to mind.
- Also write down other words associated with the sound.
- I will give you cues.

Location-based Services (LBS) and POI Discovery

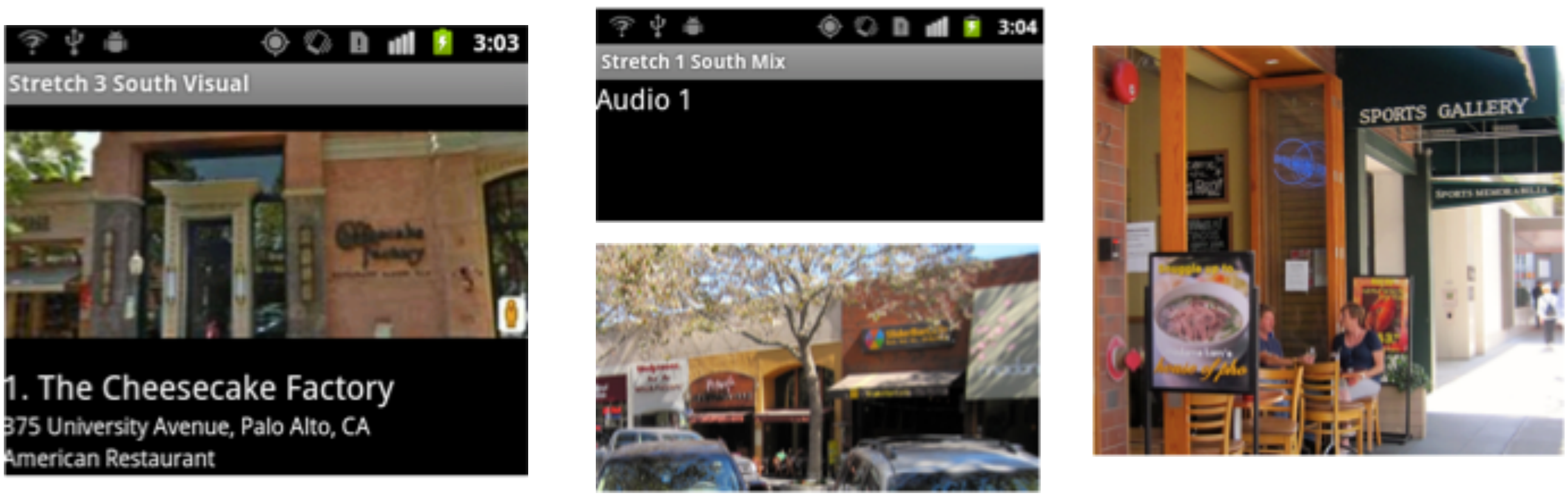
- Most Location-based services (LBS) assume that users either have a specific destination, or have a well-defined need for information.
- However, sometimes the goal of a mobile journey is to discover POIs in a pleasing, memorable and serendipitous manner.
- We present a field study comparing different kinds of cues in terms of their effectiveness (i.e. the speed, accuracy and confidence of which people identified nearby POIs) and the emotional response they elicit.
- Why a field study?

Experiment

- The experiment was conducted with 15 participants recruited from the Palo Alto area.
- The participants were randomly assigned to either the audio treatment or to be controls, such that two-thirds of the participants (10/15) experienced the audio treatment and a third (5/15) were controls.



Cue types (conditions)
Cue
Participant
Sounds
Speech
Music
Mix
Visual
Audio (treatment)
Control
University Ave.
Stretch
1
2
3
4
5
Emerson St.
Ramona St.
Bryant St.
Waverly St.
Kipling St.
Cowper St.
North



3:03
Stretch 3 South Visual
1. The Cheesecake Factory
375 University Avenue, Palo Alto, CA
American Restaurant
3:04
Stretch 1 South Mix
Audio 1
SPORTS GALLERY

Experiment Procedure

- The experiment starts on either the north or the south end of the route.
- Each cue is played for one minute.
- For each cue, participants were asked to identify the POI that the cue referred to on a paper map.
- At the end of each block, the emotional response of participants was gauged by experience sampling, using a modified PAD scale.

Our version of the PAD scale

- Pleasure:
  - Annoyed(1) - Pleased(5)
  - Bored(1) - Engaged(5)
- Arousal:
  - Calm(1) - Excited(5)
  - Relaxed(1) - Stimulated(5)
- Dominance:
  - Influenced(1) - Influential(5)
  - Guided(1) - Autonomous(5)

Analysis

- A between-subjects comparison for each block between the audio treatment and the control.
- A within subjects analysis for participants in the audio treatment to compare the outcomes of different conditions for the same users.
- For between-subjects comparison, the performance data and the emotional response data used as dependent variables were factorial and not normally distributed (thus we can not use t-tests), so we used (ordinal) logistic regression.

Analysis

- Logistic regression cannot be used on within subjects data

  - It violates the assumption about independent samples.

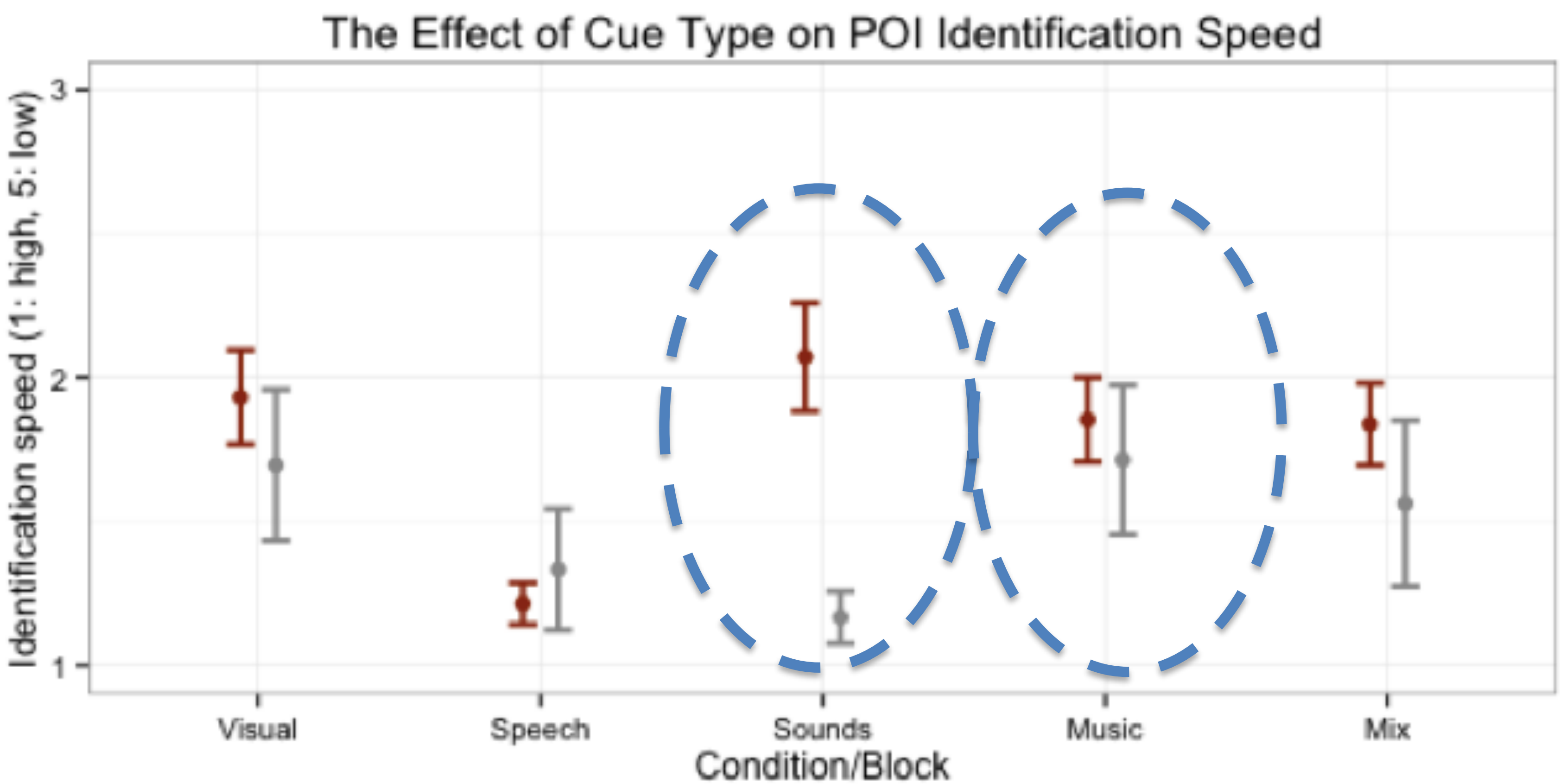
- We relied on the Aligned Rank Transform method to transform the data and subsequently applied ANOVA to detect main effects of condition on outcomes (using ARTool).
  - The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures by Wobbrock et al., from CHI2011.

Result, Effectiveness:

- For between-subjects, musicons and auditory icons were associated with greater errors (7% and 15% resp.) compared to the control participants.
- However, logistic regression comparing identification errors made with musicons and auditory icons to the errors by control participants did not reveal a main effect of condition on identification errors.
- Comparing results within subjects using an aligned rank transform followed by an ANOVA, we find a significant main effect of condition on identification.
- The contradiction here is likely due to the ANOVA analysis failing to discount some location bias.

Result: Speed

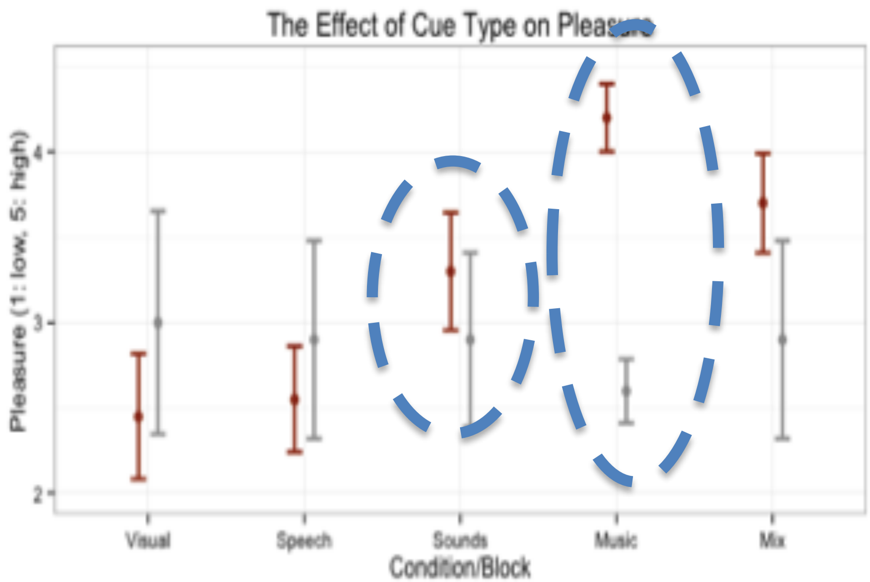
- For between-subjects, speech was associated with faster identification and auditory icons with slower identification.
- For auditory icons, an ordinal logistic regression does indeed result in a significant effect of condition on identification speed compared to controls.
- There is no significant effect for the other factors.
- For within subjects, we find that there is again a significant main effect of condition on identification speed, With auditory icons taking much longer to identify and speech cues being much quicker than the others.



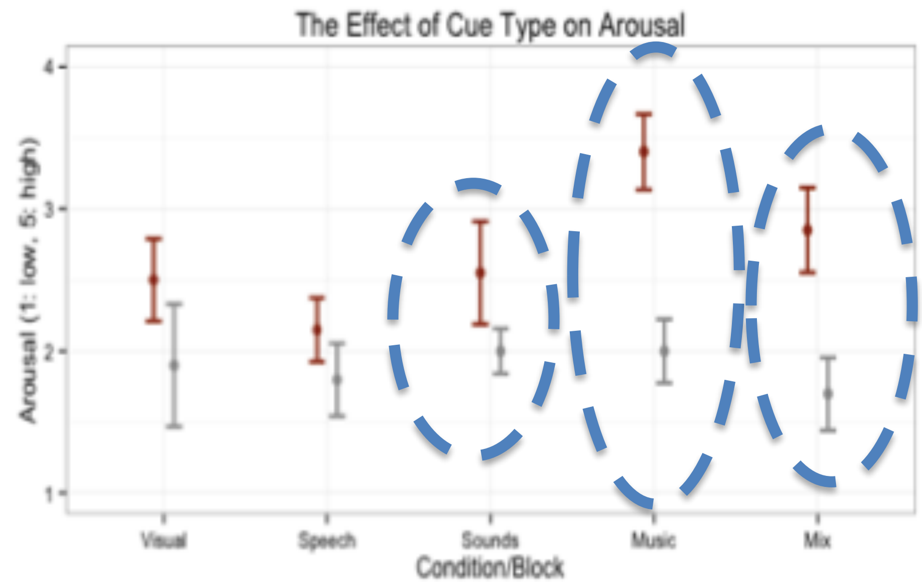
The mean reported identification speed by condition with bars showing the standard error. A speed of 1 denotes instant identification, while 5 corresponds to taking about 1 minute for identification. The maroon dots and bars (left) represent participants in the audio treatment, while the grey dots and lines (right) represent control participants for that same stretch. The Sounds condition refers to auditory icons.

Result: Pleasure and Arousal

- For musicons, an ordinal logistic regression results in a significant effect of condition on both pleasure (the pairings of Annoyed(1) - Pleased(5), and Bored(1) - Engaged(5)) and arousal (the pairings of Calm(1) - Excited(5) , and Relaxed(1) - Stimulated(5)) compared to controls.
- The results for auditory icons are not significant.
- Comparing results within subjects, we find that there is again a significant effect of condition on both pleasure and arousal, where the music condition is associated with high values and the speech condition with low values.



Mean reported pleasure by condition with bars showing the standard error. Pleasure was measured via two pairings: Annoyed(1) - Pleased(5) and Bored(1) -Engaged(5). The maroon dots and bars represent participants in the audio treatment, while the grey dots and lines represent control participants for that same stretch.



Results of arousal by condition. This graph shows the mean results of the two pairings: Calm(1) - Excited(5) and Relaxed(1) - Stimulated(2).

Summary

- Auditory icons are associated with slower and less confident POI identification compared to the visual controls.
- Musicons are only associated with lower confidence in identification, but not with slower identification.
- Auditory icons and musicons did not lead to significantly more errors in a stretch-by-stretch comparison to the visual treatment group.
- Musicons are associated with greater feelings of pleasure and arousal compared to the visual control group while auditory icons shows no difference in feelings of pleasure or arousal.

Scheduling

- We present the results of an informal survey of 15 Disk Jockeys in college radio stations
- We compile and examine the heuristics they use to create radio theme shows.
- We present an experimental study designed to evaluate various scheduling strategies for mobile music consumption.
- Using this experimental design, we present the results of a small-scale field experiment:
  - The experiment compares the user experiences of three music-feature based scheduling strategies against the gold standard of a schedule created by an experienced DJ.

DJ Techniques

- College radio DJs typically select and sequence music from a large collection of vinyl, CDs, cassettes and digital downloads to produce a show.
  - Theme shows, thus context aware.
    - “Songs of summer”,
    - “Songs of Paris”,
    - “Songs about food”…

- Most radio show slots are 1-2 hours long and consist of more than 15 songs.

DJ Techniques

- Trick to keep listeners engaged for hours: break the show up into segments of 3-4 songs or 15 minutes.
  - Breaks allows the DJs to play with the pace.
  - In some cases, breaks help to “clearing the slate”.
  - It also helps with keeping listeners’ attention.
- Within each segment, you can play with many things.
  - Genres and style, tempo, loudness, lyrics…
- DJs are encouraged to “watch their transitions”
  - Fade in, fade out…

Experiment Design

In order to compare music scheduling strategies for music on-the-go, the following requirements must be met:

- Compare scheduling of music, not music choices:
  - The experiment design must not be biased by the choice of songs in a given playlist schedule. Only the order matters.
  - We fix the selection of music tracks used in the experiment. Each condition plays exactly the same set of songs.
- Evaluate music consumption in *situ*:
  - Music in context, so a field experiment is necessary.

Experiment Design

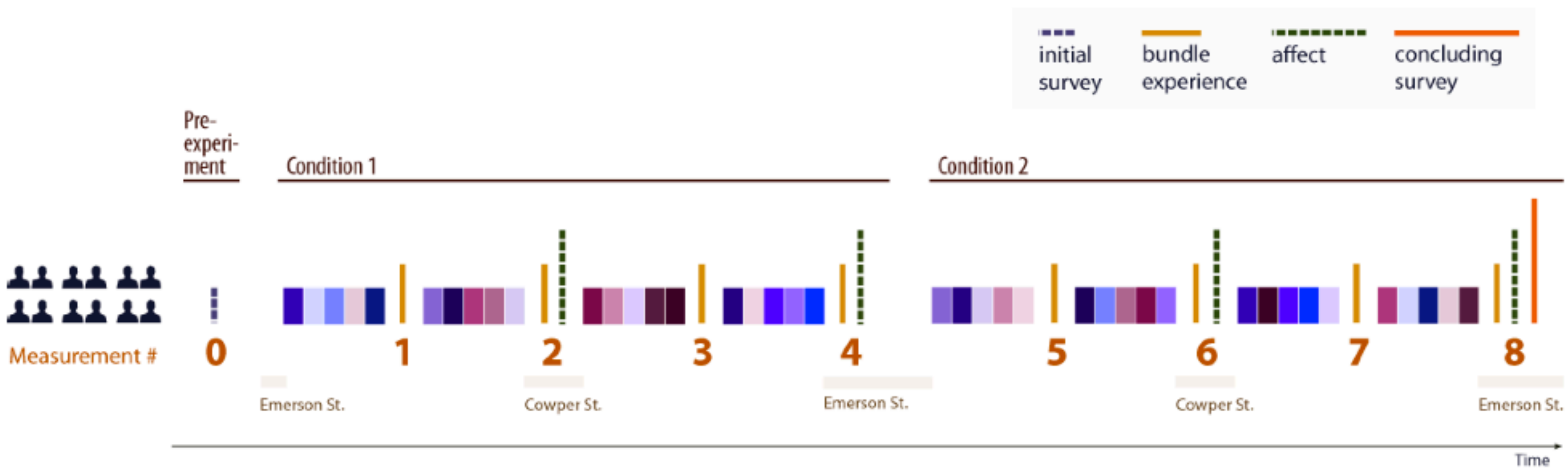
- Be independent of users' musical tastes:
  - Users' musical preferences are measured via a Web-based survey a couple of days before the field experiment.
  - Users were asked to listen to and rate a superset of the songs used in the experiment. The songs were presented in random order and include songs the users would hear eventually as part of the experiment.
- Obtain clear signals of the user experience:
  - People tend to rate music more positively during an experiment (e.g., because the user enjoyed the walk), which can obscure the differences among scheduling methods.
  - To mitigate this bias, in the data analysis process we will focus on negative ratings, which are clearer signals.

Scheduling Methods

"Experienced DJs"

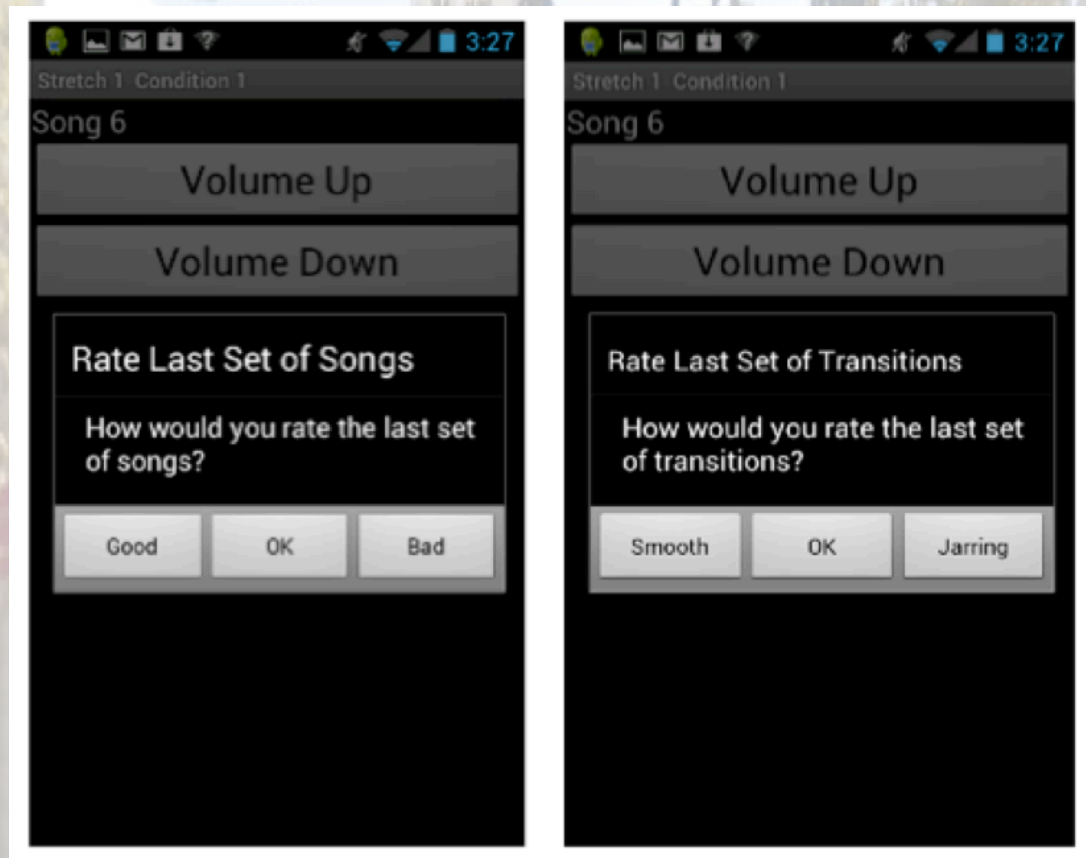
- Expert (baseline):
  - Created manually by an experienced DJ.
- Genre:
  - Relies on the genre meta-data (from *allmusic.com*) to cluster songs.
  - Each bundle contains songs of the same genre.
- Tempo:
  - Songs are ranked based on amplitude and beat, then split into bundles (see paper for a detailed formula).
- Location:
  - Assumes a set of songs with associated physical locations and orders the songs based on the expected path taken by the listener.
  - It always plays the song whose 'location' is closest to the user.

Experiment Design



- 12 participants, 48 trials, 6 participant per condition.
- Guided walk along a few blocks of a busy downtown street in Palo Alto while listening to a playlist of songs.
- Each participant experienced 2 conditions.
- The conditions are fully counterbalanced. Thus, for each condition, half of the participants experienced it as the first condition, and the other half as the second condition.

Experiment Design



- The schedule for each condition consisted of 4 bundles of 5 songs.
- We took measurements four times during the experiment, i.e. after every bundle.

Hypothesis

We tested the following hypotheses. We define μ as the (average) user experience for a particular condition:

Hypothesis 1:

- μ is greater for the Location condition than for Expert
  - The Location schedule used knowledge of the location to order music.
  - The Expert schedule used many different sources of information.

Hypothesis 2:

- μ is greater for the Tempo condition than for Genre
  - Examines the 2 conditions that used features of the songs themselves to order music.

Experiment Results

- Only negative ratings were counted.
- The measured experience, *m*, of the user for a condition is obtained for each bundle as follows:

$$m = E[m] - n_b - n_t$$

- $n_b \in \{0, 1\}$ denotes the number of negative basic ratings for a bundle during the walk,
- $n_t \in \{0, 1\}$ denotes the number of negative transition ratings,
- *E[m]* denotes the *a priori* expected experience (of 5 songs) based on the pre-experiment ratings.

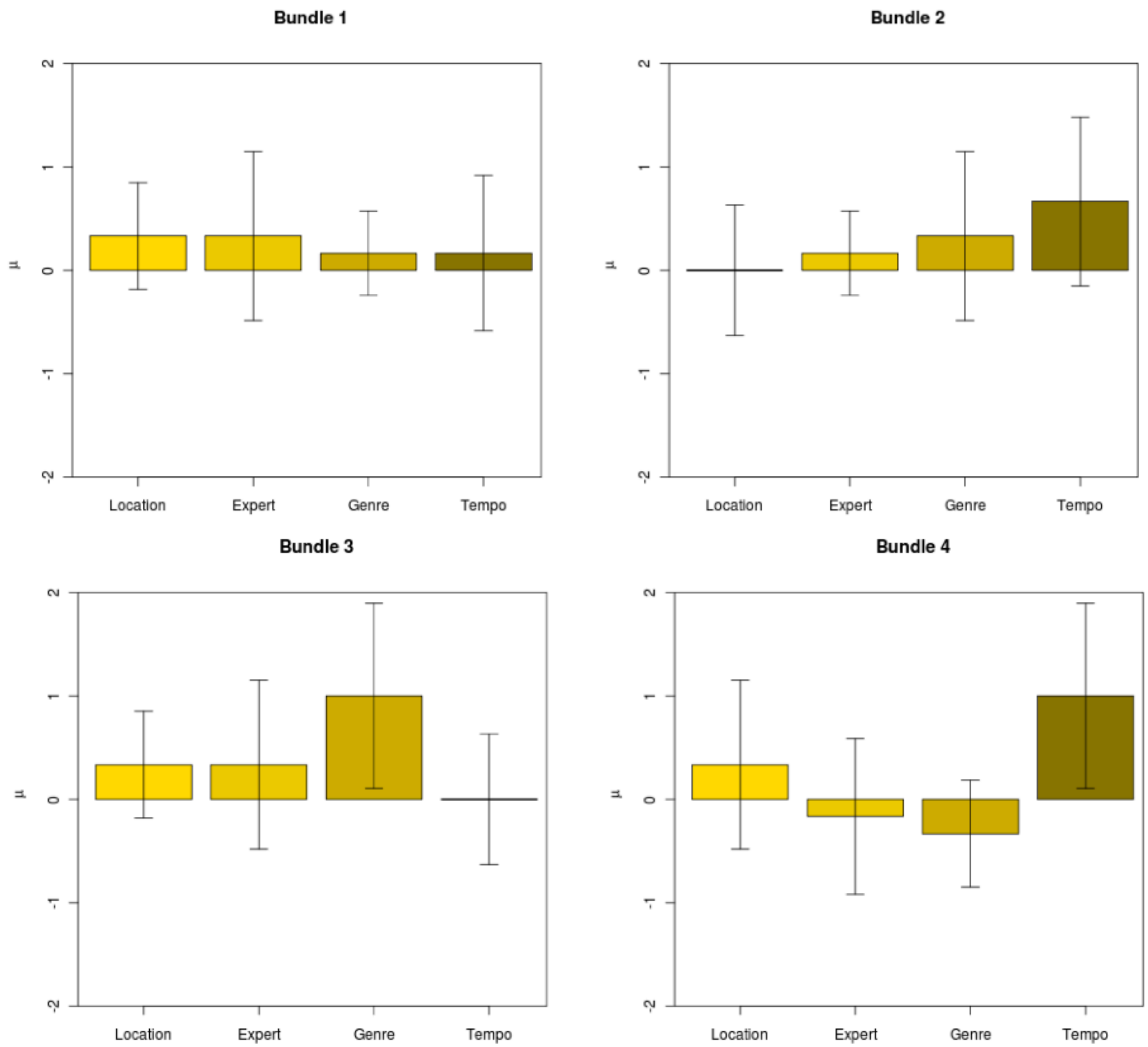
$$E[m] = \lfloor n_o/2 \rfloor$$

- Where n_o corresponds to the number of negative ratings across the 5 songs in the bundle in the pre-experiment.

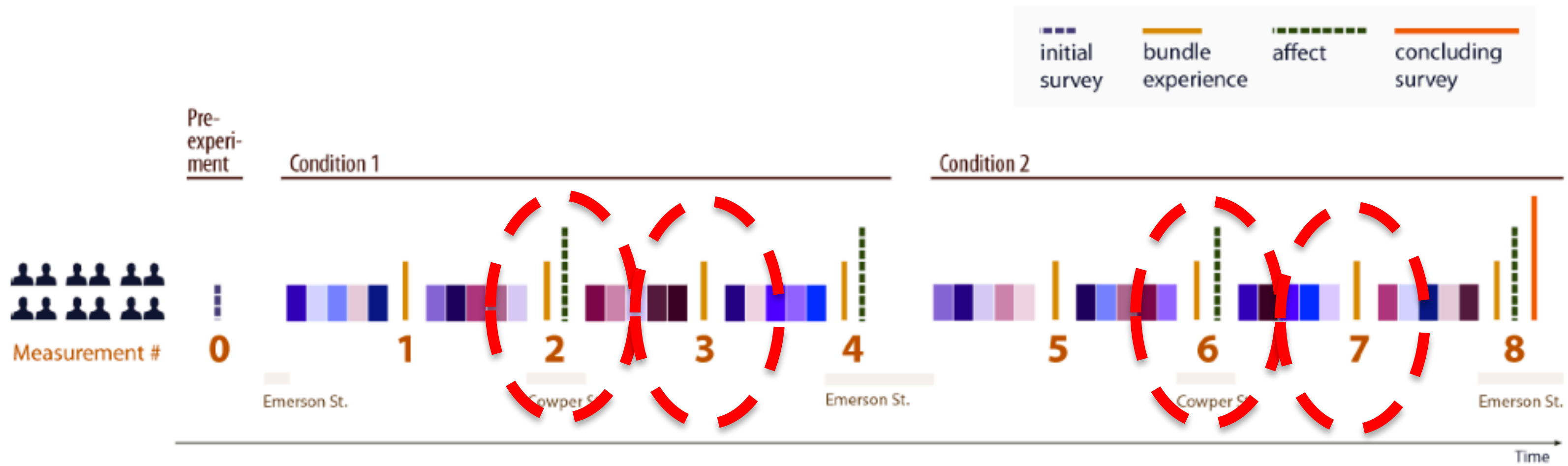
$$m \in \{-2, -1, 0, 1, 2\}$$

User Experience

The values of *m* are shown by bundle:



Experiment Results



- There were 8 samples of *m* for each user.
- However certain sample points were highly correlated for all users.
- Samples 2, 3, 6 and 7, and represent the 2nd and 3rd bundles for each user in each condition.
- By using only the samples for the 1st and 4th bundle in each condition to compare the user experience across conditions, these large correlations disappear.

Experiment Results

- After filtering, we are left with a total of 24 independent samples of *m* in the experiment: 4 samples per user and 6 samples per condition.
- For a particular condition, the user experience μ is then the average value of *m* across all 6 samples:

$$\mu_c = \frac{1}{6}\sum_{i=1}^{6} m_{i,c}$$

- We also conducted one-sided, two-sample, unpaired t-tests to determine whether the differences in means of the 6 samples per condition were significantly greater than 0.

- The values of μ_c and the p-values (Bonferroni-compensated) of the corresponding t-tests are shown here:

Hypothesis			p-value
	μ_{Loc}	μ_{Exp}	$H_0 : \mu_{Loc} \leq \mu_{Exp}$
1	.33	.08	.40
	μ_{Tem}	μ_{Gen}	$H_0 : \mu_{Tem} \leq \mu_{Gen}$
2	.58	$-.08$	.04

Conclusion 1:
We did not find significant support for Hypothesis 1, that the Location condition yields a better user experience than the Expert condition.

Conclusion 2:
We found significant support for Hypothesis 2, that the Tempo condition results in a better experience than the Genre condition.

Thank You!
(Questions?)

Results, Effectiveness:

- The contradiction here is likely due to the ANOVA analysis failing to discount some location bias. In other words, overall the controls had fewer errors but in a stretch-by-stretch and condition-by-condition comparison the differences were relatively small, making it hard to make any condition-specific conclusion.