# Impact of Item Consumption on Assessment of Recommendations in User Studies

Yeojin Joo

January 15, 2019

# Outline

- Introduction
- User Study1 : Songs
  - Method
  - Results and Discussion
- User Study2 : Movies
  - Method
  - Results and Discussion
- Conclusions and Outlook

# Introduction

- Measuring user experience becomes important in Recommender Systems research.
- Recommended products are usually represented by textual descriptions, pictures and metadata.
- Only in rare cases, it is possible to actually consume them.
- We conducted two user studies to investigate pre- and post-consumption assessments of recommendation quality and aspects.

# User Study1 : Songs

## Method

- We hypothesized that actually listening to recommended songs makes a difference in assessment.
- Set two conditions.
  - S1 : questionnaires Pre and Post consumption ($N_{S1} = 21$)
  - S2 : questionnaires only Post ($N_{S2} = 19$)
- 5 recommendations with song titles, artist, album titles and covers displayed.
- All items were assessed on a 1-5 Likert-scale.
- Linear mixed-effect model
  - fixed factor : condition($S_1, S_2$), point in time (Pre, Post)
  - specified point in time as a repeated measurement

# User Study1 : Songs

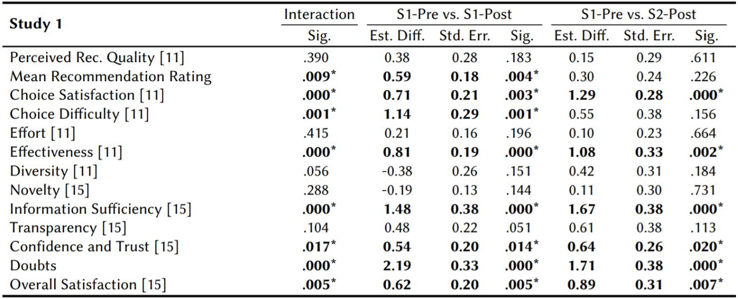
## Results and Discussion

- S1-Pre vs S1-Post (within-subject)
    - The difference of mean rec. rating is larger, the lower perceived rec. quality in S1-Pre. (r=-.709, p<.000)
    - The ratings are nomally distriuted with less variance (more strong opinion) in S1-Post(0.33) than S1-Pre(0.77).
    - The difference between S1-Pre and -Post is higher, the fewer items are known.(r=-.492, p=.023)
    - Participants were more satisfied when they found information sufficient in S1-Pre. (r=.745, p <.000)

# User Study1 : Songs

## Results and Discussion

- S1-Pre vs S2-Post (between-subject)
  - Experiencing the songs led to higher perceived information sufficiency and fewer doubts.
  - More satisfied, the higher information sufficiency. (r=.626, p=.004)



| **Study 1** | Interaction | S1-Pre vs. S1-Post | | | S1-Pre vs. S2-Post | | |
|---|---|---|---|---|---|---|---|
| | Sig. | Est. Diff. | Std. Err. | Sig. | Est. Diff. | Std. Err. | Sig. |
| Perceived Rec. Quality [11] | .390 | 0.38 | 0.28 | .183 | 0.15 | 0.29 | .611 |
| Mean Recommendation Rating | **.009$^*$** | **0.59** | **0.18** | **.004$^*$** | 0.30 | 0.24 | .226 |
| Choice Satisfaction [11] | **.000$^*$** | **0.71** | **0.21** | **.003$^*$** | **1.29** | **0.28** | **.000$^*$** |
| Choice Difficulty [11] | **.001$^*$** | **1.14** | **0.29** | **.001$^*$** | 0.55 | 0.38 | .156 |
| Effort [11] | .415 | 0.21 | 0.16 | .196 | 0.10 | 0.23 | .664 |
| Effectiveness [11] | **.000$^*$** | **0.81** | **0.19** | **.000$^*$** | **1.08** | **0.33** | **.002$^*$** |
| Diversity [11] | .056 | -0.38 | 0.26 | .151 | 0.42 | 0.31 | .184 |
| Novelty [15] | .288 | -0.19 | 0.13 | .144 | 0.11 | 0.30 | .731 |
| Information Sufficiency [15] | **.000$^*$** | **1.48** | **0.38** | **.000$^*$** | **1.67** | **0.38** | **.000$^*$** |
| Transparency [15] | .104 | 0.48 | 0.22 | .051 | 0.61 | 0.38 | .113 |
| Confidence and Trust [15] | **.017$^*$** | **0.54** | **0.20** | **.014$^*$** | **0.64** | **0.26** | **.020$^*$** |
| Doubts | **.000$^*$** | **2.19** | **0.33** | **.000$^*$** | **1.71** | **0.38** | **.000$^*$** |
| Overall Satisfaction [15] | **.005$^*$** | **0.62** | **0.20** | **.005$^*$** | **0.89** | **0.31** | **.007$^*$** |

# User Study2 : Movies

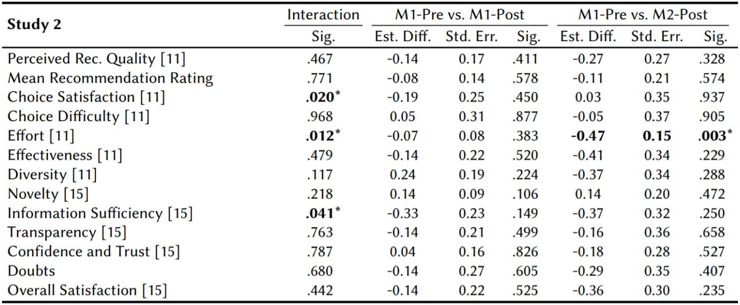
## Method

- Designed similar to study 1.
- Set two conditions.
  - M1 : questionnaires Pre and Post consumption ($N_{M1} = 21$)
  - M2 : questionnaires only Post ($N_{M2} = 19$)
- 3 recommendations with movie titles, genres, posters, metadata on director and cast, and description texts by the article's author.

# User Study2 : Movies

## Results and Discussion

- This result is clearly in contrast to study 1.
- It seems participants were able to accurately estimate whether they will like recommended items.



| **Study 2** | Interaction | M1-Pre vs. M1-Post | | | M1-Pre vs. M2-Post | | |
|---|---|---|---|---|---|---|---|
| | Sig. | Est. Diff. | Std. Err. | Sig. | Est. Diff. | Std. Err. | Sig. |
| Perceived Rec. Quality [11] | .467 | -0.14 | 0.17 | .411 | -0.27 | 0.27 | .328 |
| Mean Recommendation Rating | .771 | -0.08 | 0.14 | .578 | -0.11 | 0.21 | .574 |
| Choice Satisfaction [11] | **.020**$^*$ | -0.19 | 0.25 | .450 | 0.03 | 0.35 | .937 |
| Choice Difficulty [11] | .968 | 0.05 | 0.31 | .877 | -0.05 | 0.37 | .905 |
| Effort [11] | **.012**$^*$ | -0.07 | 0.08 | .383 | **-0.47** | **0.15** | **.003**$^*$ |
| Effectiveness [11] | .479 | -0.14 | 0.22 | .520 | -0.41 | 0.34 | .229 |
| Diversity [11] | .117 | 0.24 | 0.19 | .224 | -0.37 | 0.34 | .288 |
| Novelty [15] | .218 | 0.14 | 0.09 | .106 | 0.14 | 0.20 | .472 |
| Information Sufficiency [15] | **.041**$^*$ | -0.33 | 0.23 | .149 | -0.37 | 0.32 | .250 |
| Transparency [15] | .763 | -0.14 | 0.21 | .499 | -0.16 | 0.36 | .658 |
| Confidence and Trust [15] | .787 | 0.04 | 0.16 | .826 | -0.18 | 0.28 | .527 |
| Doubts | .680 | -0.14 | 0.27 | .605 | -0.29 | 0.35 | .407 |
| Overall Satisfaction [15] | .442 | -0.14 | 0.22 | .525 | -0.36 | 0.30 | .235 |

# Conclusions and Outlook

- Participants in some cases cannot adequately assess all aspects of RS, especially those related to user experience.
- Assessment highly depends on domain as well as type and amount of information provided alongside recommendations.
- We suggest to avoid comparisons across different settings and to pay attention in user experiments without consumption.