# Homework 1

Measurement and Evaluation of HCC Systems

## How to hand in this homework

- Please email the homework to me as a PDF.
- Late assignments get a penalty of 20% when submitted after the deadline, plus an additional 10% per hour late.
- Make sure you include the R input you used to get to your answer, but do not "dump" the resulting R output on the paper. Copy from the output selectively, and explain it in your own words.
- You may collaborate on this homework, but not copy from others… again, please write your answers in your own words.
- Please include a collaboration statement that says: "I collaborated on this homework with [name]." or "I worked alone on this homework"

## Dataset

You are going to use the dataset from an experiment conducted on the TasteWeights recommender system. The TasteWeights system uses the overlap between your Facebook "likes" and the "likes" of your Facebook friends to give you music recommendations. This works as follows:

- Your friends are given a "weight" based on the overlap between your music likes and your friends' music likes: a friend with 2 overlapping likes gets a weight of 2.
- Then, the friends' *other* music likes—the ones that are not among are user's likes—are tallied by weight: if three friends with weights 2, 5 and 7 all like Coldplay, then Coldplay gets a score of 2+5+7 = 14.
- This tallied list is then sorted by score, and the Top 10 is presented to the user.

### *Manipulations*

We tested two features of the TasteWeights system. These are translated into two manipulations: "Control" (3 levels) and "Inspectability" (2 levels). All combinations are tested, so there are 6 conditions. This is a between-subjects experiment.

- **Control (see Figure 1)**
  - **None:** The user has no control; the system calculates the recommendation with the default method (see above).
  - **Item:** In this condition, participants can weigh their likes before the system calculates their recommendations. The friend weights are now the sum of the weights of the overlapping items (rather than just the number of overlapping items).
  - **Friend:** in this condition, participants can adjust the friend weights before the system calculates their recommendations.
- **Inspectability (see Figure 2)**
  - **List view:** The recommendations are presented as a sorted list.
  - **Graph view:** The recommendations are presented as a graph that shows their music likes, their friends, and their recommendations. By hovering over the lists, they can see how these things are connected to each other.

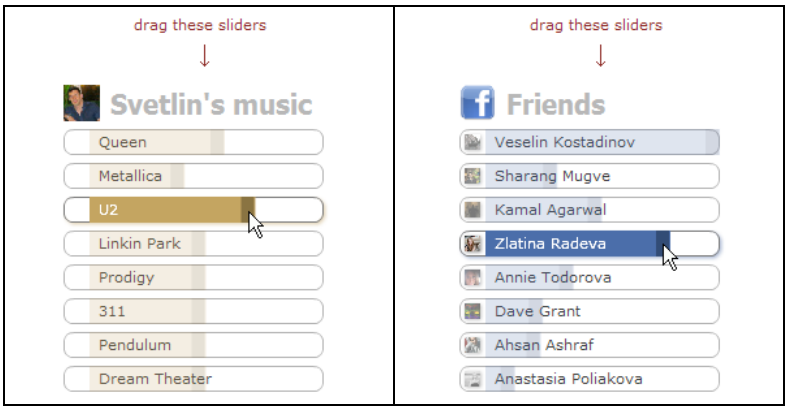


*Figure 1: Item control (left) and Friend control (right)*


*Figure 2: List view (left) and Graph view (right)*

### *Procedure and measurements*

At the start of the experiment, the participants get a questionnaire measuring the following personal characteristics:

- **Expertise:** Participants' music expertise, measured with four 5-point scale questions (ranging from –2 to +2), translated into a sum score.
- **Trust:** Participants' trusting propensity, measured with three summed 5-point scale questions.
- **Familiarity:** Participants' familiarity with recommender systems, measured with two summed 5-point scale questions.

Participants are then taken to the experiment where they are asked to do the following:

- Control the items/friends (or not), depending on the condition they are in
- Inspect the recommendations (either in graph or list, depending on the condition), the time they take to inspect the recommendation is captured in the variable **time** (measured in seconds)
- Rate the recommendations on a 5-star rating scale (these ratings are averaged as the variable **rating**)
- Indicate which of the recommendations they already knew (the number of known recommendations is captured as the variable **known**)

At the end of the experiment, the participants get another questionnaire measuring the following subjective evaluations:

- **Satisfaction:** Participants' overall satisfaction with the TasteWeights system, measured with seven summed 5-point scale questions.
- **Quality:** Participants' perception of the overall quality of the recommendations, measured with six summed 5-point scale questions.
- **Perceived_control:** Participants' perception of their control over the recommendations, measured with four summed 5-point scale questions.
- **understandability:** Participants' understanding of how the recommendations were calculated, measured with three summed 5-point scale questions.

If you want to learn more about this experiment, you can read Knijnenburg et al. (2013) "Inspectability and control in social recommenders". DOI: 10.1145/2365952.2365966

## Question 1. Describing data

a. Use R to get the mean and standard deviation of the "rating" variable.
b. Use this value to calculate a 95% confidence interval for the mean rating.
c. Explain what this confidence interval means.

## Question 2. Correlation between rating and perceived quality

To test its validity, we want to make sure that the subjective measure "quality" is correlated with "rating".

a. What kind of validity are we testing this way?

Let's first test the assumption of normality.

b. Plot a histogram of the "rating" variable.
c. Use R to get the skewness, kurtosis and Shapiro-Wilk test for the "rating" variable. Are the skewness and kurtosis values significant?
d. Are these values and/or the histogram cause for concern? Why (not)?
e. Repeat questions b-d for the "quality" variable.

Now let's move to the correlation itself.

f. Create a scatterplot for quality (X) and rating (Y), and put a linear trend line on the plot.
g. Give the correlation between quality and rating, and its p-value. Use the correct correlation method, given your conclusions about the normality of "quality" and "rating".
h. What can we conclude, given these values? How much of the variance in "rating" is explained by "quality"?

## Question 3. Regression of satisfaction

Our final goal is to explain what makes people satisfied with the system. We believe that it is mainly due to the quality of the recommendations, but perceived control and understandability may also have an effect.

Let's start with a power analysis.

a. We want to be able to detect $R^2$ increases of at least .06. What power do we have to detect such an effect with $\alpha$ = .05? (hint: in G*Power, use t-test, linear multiple regression; go for a 2-tailed test)
b. How many participants would have been enough to achieve a power of 0.85?

Now let's try some regressions.

c. Run model1: a simple regression of satisfaction (Y) on quality (X). Interpret the results. This includes the overall model fit ($R^2$, F-test, significance) well as the model parameter for quality (slope b and its significance).
d. Can we make causal statements about the results? Why (not)?
e. Run model2: Add perceived control and understandability to the regression. Interpret the results ($R^2$, its increase, and the b-parameters of the predictors and their significance).
f. Test whether model2 improved upon model1. Report the results of this test (the F-test and its significance).

That looks good. Now let's see if we should add any other variables.

g. Use correlation tests to find out whether you should add rating, known, and/or time to the model.
h. Run model3: Add the variables that you think should be added. Interpret the results ($R^2$, its increase, and the b-parameters of the predictors and their significance).
i. Test whether model3 improved upon model2. Report the results of this test (the F-test and its significance).
j. Why do the regression outcomes for the added variables differ from the correlation outcomes?
k. Create a nice table of the 3 models (like Field's Table 7.2), and write a short conclusion about your findings.

Finally, let's test our assumptions.

l. Use the methods from question 2b-d to test the normality of the satisfaction variable.
m. Should we also test the normality of the predictor variables in our model? Why (not)?
n. Find the data points with large standardized residuals (>1.96 or < -1.96) based on model2. Which data points may be of concern? Do you have an excessive number of outliers? What about substantial outliers? Extreme outliers?
o. Get the Cook's distances, leverages, and covariance ratios for the data points with large standardized residuals. Interpret the results. Are you still concerned?
p. Test the variance inflation factors (VIFs). Are they problematic?
q. Finally, plot the model residuals to check for normality, homoscedasticity, and linearity. Are any of these assumptions violated, based on the plots?

Since satisfaction was not nicely normally distributed, we should do a bootstrapped regression.

r. Run model2 as a bootstrapped regression with 2000 repetitions. Interpret the results.

## Question 4. Regression of understandability

We believe that the inspectability manipulation should have an effect on understandability. Specifically, we believe that the graph view *increases* users' understandability compared to the list view.

a. What is the experimental hypothesis?
b. What is the null hypothesis?

Let's run a regression to test the hypothesis.

c. Create a contrast: graph_v_list
d. Run the regression of understandability (Y) on inspectability (X). Interpret the results (b-parameter, significance). Make sure you report the one-sided p-value!
e. Can we make causal statements about the results? Why (not)?

f. Test for heteroscedasticity with a boxplot, Levene's test, and the variance ratio test. Is there a cause for concern?
g. Given our experimental hypothesis, what would we be able to conclude if the t-value had been *negative* (i.e., the graph was *less* understandable than the list)?

## Question 5. Regression of perceived control

We believe that the control manipulation should have an effect on perceived control. Specifically, we believe that having friend or item control *increases* users' perceived control compared to "no control" condition.

a. Create two contrasts: item_v_none and friend_v_none
b. Run the regression of perceived_control (Y) on control (X). Interpret the results (b-parameter, significance). Make sure you report the one-sided p-values!

What if we were interested in the difference in perceived control between the item control condition and friend control condition?

c. Think of a way to test this. Report on the results.

Let's say that we had 5 control conditions instead of 3.

d. What would be the problem with testing all possible differences between conditions?