STATS 551 Final Project Proposal
Yuyan Han & Qinggang Yu

Goal

People often do not have accurate perception of their self-knowledge. Prior work has documented the *Dunning-Kruger effect*, which suggests that across multiple domains, bottom-performers overestimate their performance, while top-performers tend to underestimate their performance. Such an effect also happens when people judge their performance relative to others: bottom-performers overestimate their ranking in terms of performance, yet top-performers underestimate.

So far, it remains unclear what contributes to this miscalibration in social comparison. In a typical survey studying the effect, respondents are asked to pick the best answers for a set of multiple-choice questions, rate their confidence in each pick, and/or give an overall estimation of their own performance and rank their performance relative to others in percentile. Oftentimes we do not have information on how respondents make specific estimations about others' performance, and how they use such estimations to rank themselves relative to others. This lack of information makes it difficult to separate the individual mechanisms that contribute to the biased ranking in social comparison.

The goal of the present project is to examine one possible mechanism through which respondents might integrate their beliefs about the self and their beliefs about others to reach a final ranking of their ability using a Bayesian model. In our model of self-assessment in social comparison, we assume that people's Bayesian inferences about their ranking are based on: (1) prior beliefs about their ability relative to others, (2) confidence on their choices for specific questions, and (3) their beliefs about others' answers to the questions. The present project aims to test whether adding in a parameter that accounts for the perceived similarity between the self and others will improve the model fit. We hypothesize that given one's confidence on a particular item, the more similarities one perceived between the self and the others, the more likely they will believe that others chose the same answer as they did. Furthermore, this perceived similarity might increase with respondents' true ability due to a well-known cognitive bias called the *curse of knowledge* or *curse of expertise* (i.e. an individual falsely assumed that others' shared the same expertise they themselves have during communication) since top performers presumably have higher chance to display the bias. We believed that the perceived similarity might play a critical role in the miscalibration during social comparison.

Data

The data used in this project are collected by Yuyan Han and Dr. David Dunning through Amazon Mechanical Turk in the year of 2017 and 2018. The data are composed of three separate datasets examining the Dunning-Kruger Effect through surveys of the same format, but on different topics, namely health care policy, abstract reasoning, and global literacy.

The health care policy survey was composed of 18 two-alternative multiple-choice questions on the topic of Affordable Care Act. For each question, participants (1) picked the best choice for the question as they believed; (2) rated their beliefs, in percentage, that each choice was the correct answer; (3) estimated the percentage of their peer respondents that would pick each choice as the best answer. The probabilities assigned to each choice of a question summed up to 1 for both (2) and (3). Here is a sample question:

> "New employees must be offered insurance within____ days of their start date.
> A. 60 days; B. 90 days.
> What do you think is the probability of each choice being the correct answer?
> What percentage of MTurk workers taking this survey do you think will choose each choice as the correct answer?"

At the end of the survey, participants ranked their performance compared to others in terms of percentile.

The abstract reasoning survey consisted of 12 four-alternative multiple-choice questions adapted from the Raven's Advanced Progressive Matrix, and the global literacy survey consisted of 16 four-alternative multiple-choice questions on global geography, international systems etc. The formats of these surveys were exactly the same as the health care policy survey.

The final sample size was 80 for the health care policy dataset, 197 for the abstract reasoning dataset, and 184 for the global literacy dataset, after excluding inattentional participants.

The datasets provide all the information we need for our model, that is, respondents' confidence for each question, their beliefs about other people's answers, their true ability based on the performance score, and their final percentile ranking in social comparison. We will not account for the data collecting process in our analysis though.

However, there are several deficits with our datasets. Frist, all the participants were MTurk workers. These workers volunteered to take our paid surveys, and they could leave the survey at any time. Therefore, there might be a self-generated selection bias even though we attempted to get a sample as random as possible. Second, these workers are not a nationally representative sample. Third, the sample sizes are small. These factors limit our ability to draw a reliable and externally-valid conclusion about the general population. Had we had an opportunity to recollect the data, we would increase the sample size, and reach out to more platforms. Nevertheless, preliminary analysis of the data showed that they were consistent with previous findings on calibration and the Dunning-Kruger Effect, and we had three separate datasets that allow a check of replicability. This increases our confidence in finding some meaningful results for the current project.

Plan for collaboration

The data was collected by Yuyan for a related project on the Dunning-Kruger Effect. Both authors contributed to the framing and the analytic plans of the project. The authors will independently analyze the data according to the analytic plan, and both will contribute to the presentation and the writing of the final report. The authors will meet twice a week to ensure smooth and timely progress of the project.

Plan for data analysis

We will adopt the item response theory (IRT) to approach this question. We assume that the probability that a respondent believes that another person will choose the same answer for an item, denoted as Y, is a mathematical function of the person and items. In this case, the person trait refers to the respondent's beliefs about their ranking relative to others, denoted as θ_p, and a general perceived similarity between themselves and others, denoted as α; the item trait refers to respondent's confidence on specific items, denoted as X, and respondent's level of performance Z. That is, $P(Y \mid \theta_p; \alpha, X, Z) = f(\theta_p, \alpha, X, Z)$. Therefore, respondents' posterior beliefs about their ranking relative to others θ_p, after a given question, is determined by their prior beliefs about their ranking, their perceived similarity with others, and their confidence on their choice, *X,* their performance *Z*, and the proportion of people they believe would choose the same answer, *Y*, would be:

$$P(\theta_p \mid \alpha, X, Z, Y) \propto P(\theta_p) * P(Y \mid \theta_p; \alpha, X, Z)$$

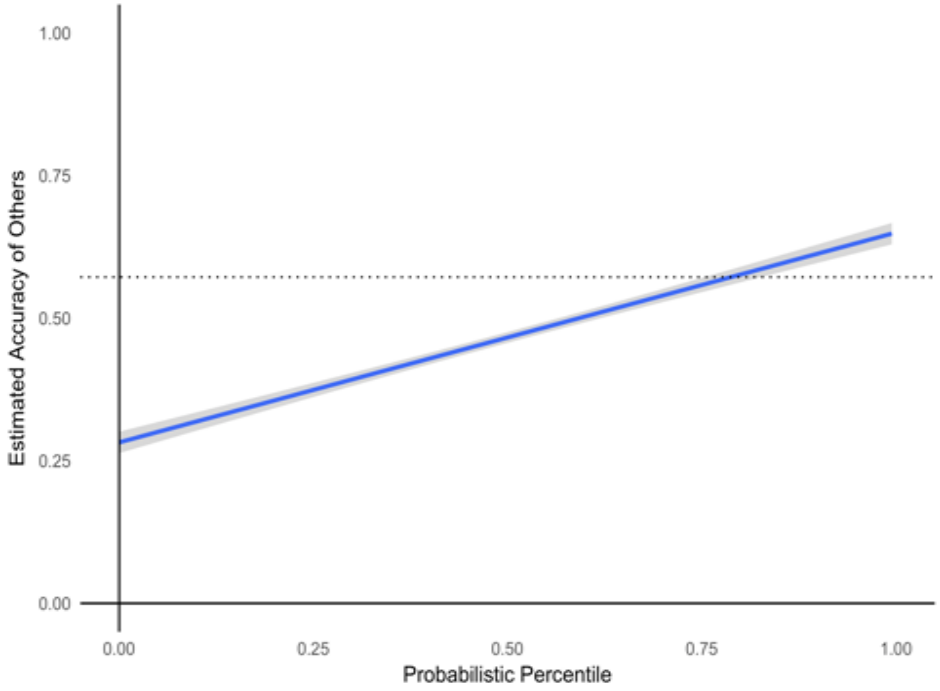
We will use Monte Carlo simulation to calculate the posterior distribution of θ_p for each level of performance Z, and then compare expected value of θ_p to the actual data of ranking. We will adjust α to see which value of α, (i.e. what level of similarity) will produce the highest model fit, and whether α instead of being a constant, should vary with performance level.

One difficulty we encountered is how to specify the mathematical model of the sampling distribution. We will consult with the GSI and aim to fit the preliminary during the week of March 15. We will run simulations and adjust the model accordingly in the two weeks that follow. We will use the remaining two weeks for the writing and the presentation.

Preliminary data analysis

(1) All three datasets showed a large Dunning-Kruger Effect in social comparison as shown by the linear regression model of perceived percentile ranking on the actual performance percentile ranking. The slopes of the calibration line were respectively 0.24 (95% CI: [0.10,0.38]), 0.22 (95% CI: [0.08,0.36]), and 0.36 (95% CI: [0.24,0.47]), far from a perfect calibration.
(2) Participants seemed to assume similarities between their own answers and their peers. Respectively for the three datasets, people perceived 64%, 60%, and 56% of other Mturkers would choose the same answer as they did while a chance-level assumption would be 50%, 25%, and 25% respectively.

(3) The perceived similarity was also found to be larger among top performers than bottom performers. On item level, a linear regression model showed that the perceived proportion of others choosing the same answer increases with performance level (β =0.181, p<0.001). On survey level, top performers tended to overestimate others' performance and bottom performers tended to underestimate others' performance (β =0.368, p<0.001, as shown in the figure below).



These preliminary results seem to be consistent with our hypotheses about the perceived similarity between the self and others, thus increasing the hope for our hypotheses.