How Videos of People Smiling Can Explore the Psychology of Judgment and Decision Making

James Kajdasz, Kyle Eschen

August 7, 2018

Our project aimed to explore two findings related to human judgment and decision making by analyzing a large data set of online survey data. The first finding relates to the "wisdom of the crowd" phenomenon. When trying to answer some question of uncertainty, it's been found that if you take a large number of individual estimates and then average them, the average is often very accurate. This phenomenon was popularized in James Surowiecki's book The Wisdom of Crowds. A second finding is what's known as the overconfidence bias. People tend to be confident more than they are correct. Such overconfidence often contributes to decision aide neglect (Sieck & Arkes, 2005), making it a pivotal topic for decision scientists who may find themselves data modeling to make decision aides. We demonstrate these principles and perhaps advance the field by examining a data set of individuals attempting to guess whether a person smiling is being genuine or not.

Data Description

In the course of his research, Dr Paul Ekman created videos of people smiling. Sometimes the smile was a genuine smile that occurred spontaneously while participants watched video clips of baby animals playing. Sometimes the participants were simply asked to smile. There are physical differences between the two that the observant may notice. French neurologist Guillaume Duchenne wrote of these differences in the early 19th century. A genuine smile (also known as a Duchenne smile, in honor of the French neurologist) is characterized by engaged muscles around the eyes along with a raising of the corners of the mouth. These muscles are difficult for people to operate consciously, and so a smile that is not genuine can lack these characteristics, or look so forced as to appear insincere. For a time, these smile videos could be viewed on the BBC website, and internet browsers could test their ability to tell the difference. As part of an earlier research project, one of the authors of this paper used these smile videos and incorporated them into an internet survey on SurveyMonkey that people can still take today: https://www.surveymonkey.com/r/SmileRead

In the survey, a few demographic questions are asked and then the participant views 20 videos of people smiling. Data collected includes:

Start Time/ End Time: (date and time survey was started and exited)

IP Address: For example: 68.102.240.36

Confidence: Participant asked how many (of 20 videos) will you get correct?

Demographic data: Gender, Age, Height, # children

Job experience: Years worked in a job where you had to 'read' people

Data Preparation

After the original research project was completed, the SurveyMonkey smile survey was left open. Without any advertisement, people continued to find the survey and take it. For three years, the number of respondents continued (and continues) to grow. At the time of our analysis, a total of 94,577 people had accessed the survey. We attempted to download this database through the SurveyMonkey website, but were unable to. Apparently the database was too big and the servers were timing out before we could receive the data. After

10 days interacting with SurveyMonkey technical support, they were finally able to do a manual pull of the data from their location and send us a .csv file of the data. The total file memory was 22.4MB.

The raw .csv file data was relatively clean, but needed to be scrubbed and reformatted for our purposes. Pertinent variables were brought into a pandas dataframe with named columns. Next, people who did not complete the entire survey were eliminated from our data. This was done by examining the last question, and identifying people who did not submit an answer. If they answered the last question, then they also answered all preceding questions, as the survey will not proceed to the next question unless an answer to a current question is submitted. Once incomplete surveys were removed, the sample size dropped to 64,427 for a completion rate of over 68%. The high completion rate is likely a reflection of the relatively enjoyable nature of the survey. Our data was now ready to be applied to our research questions.

Wisdom of the Crowds

In 1907, Sir Francis Galton attended a weight-judging competition at the West of England Fat Stock and Poultry Exhibition. Competitors bought some 800 tickets on which they guessed the weight of an ox. Galton collected the tickets and was interested to note that the mean of all the guesses (1,197 lbs) was within 1 pound of the actual weight of 1,198 lbs (Galton 1907, Galton 1907b). The purpose of Galton's original article was to promote the use of the median as a better representation than the mean for central tendency, but this fact is overlooked in the popular re-telling of the story. This phenomenon, where the aggregate of many individual estimates tends to provide a very accurate estimate, has been demonstrated in a wide variety of fields to include estimates of magazine sales (Ashton & Ashton, 1985), airline passenger forecasts (Bates & Granger, 1969), movie box-office earnings, mobile phone usage, and pop-music chart positions (Spann, Martin & Skiera (2003). The phenomenon has been termed "The Wisdom of the Crowd", and was popularized by the book written by James Surowiecki of the same name (2005).

Past research demonstrated and replicated the principles of the "Wisdom of the Crowd" with the smile task data. When performance across all smiles was evaluated, individuals achieved 68% accuracy (13.6/20 smiles). When theoretical groups were created randomly, it was found that overall performance increased with group size (Kajdasz, 2014). Figure 1 represents the total average performance on all 20 videos with theoretical groups of various sizes. However, this line only represents a single task with, in this case, an average individual performance rate of 68%. For tasks of higher or lower difficulty, one can imagine additional lines fitting on this graph, representing tasks of various difficulty levels, such as in the hypothetical graph in Figure 2.

The original analysis was accomplished largely manually, with much less data (n = 217), utilizing SPSS. Before taking W200, it was not possible to create the hypothetical graph in Figure 2. However, after completing the W200 Python class, we now have the ability to obtain a deeper insight for the first time with a much larger dataset. We will do this by treating each smile video as an individual task. Each smile video has its own individual performance rate. Some videos are obvious fakes/obviously genuine. Some are very difficult to determine. We can use each of the 20 videos to represent a task of varying difficulty. The accuracy rate, across all 64,427 individuals, is offered in Figure 3.

Our procedure to collect the necessary data to make Figure 2 is now described. Let's say we are collecting data for a nominal group of five individuals. The term 'nominal group' is used here to denote that this is a group in name only-the individuals never actually meet. To create our nominal group, five individuals were selected at random from the total sample of 62,427. These five individuals then had their answers compared for a single smile video. If the majority of the group assessed the video correctly, then the group was assessed as returning a 'correct' answer. This process was repeated another 30 times so that a single sample represented 30 groups of 5. The average performance of this sample was recorded as a single data point. This sampling process was repeated another 30 times, so that estimates of variance could be obtained. The final data is a list of 30 numbers that represents the performance of each sample, where each sample represents 30 groups.

The sampling process above was repeated for every smile video, and for every possible group size from 1 to 99 (odd-sized groups only to eliminate the possibility of a tie), with new individuals being randomly chosen with replacement every time. The maximum sample selected, 30 groups of 99 (for a total of ~3000), still represents

less than 5% of the total sample of 62,427, assuring that the samples can be considered independent and identically distributed according to general rules of thumb. (Devore, 2016, p222).

The resulting data from this process is a Python Pandas data frame with three variables: Difficulty: This is the difficulty of the task of a single smile video. The difficulty is derived as the individual accuracy rate across all individuals in the sample (see Figure 3). Group Size: Group sizes varied from a group of one, all the way up to a group of 99. Only odd-numbered groups were calculated to avoid ties. Sample Performance: Each cell represents the average performance of 30 groups of size X.

Once all the group simulations were run, the next step was to plot the data and attempt to create the hypothesized graph in Figure 2. Plotting all 20 lines (one line for each video) would make for a muddy graph. Instead, we chose 5 lines that represented the broadest range of task difficulty: Question 17 (55% accuracy), Q20 (63%), Q12 (75%), Q19 (85%) and Q18 (94%). A Python script was used to slice the appropriate columns and plot them with MatPlotLib. The results are included as Figure 4.

Our data replicated the 'Wisdom of the Crowd' phenomenon. Aggregated groups of individuals do substantially better than individuals. Further, we have simulated data for groups of various sizes across a variety of task difficulties. This last point is significant. The simulated data can be fitted to a model where:

$$PredictedGroupPerformance = \beta_0 + \beta_1(TaskDifficulty) + \beta_2(GroupSize)$$

This model can be used in a number of interesting ways.

Predicting Group Performance: The most obvious application is the model can be used to predict group performance for a task of some assessed difficulty and group size.

Optimizing Group Size: The model can be used to optimize the proper size of a group so that their performance matches some desired level after individual difficulty of the task is estimated.

Benchmarking Other Group Techniques: There are other ways for groups to work together besides averaging their individual input. For example, Rosenburg (2017) applies a "swarming technique" to organize group input to perform the smile video task. The model in this paper can be used as a benchmark so that other techniques can be compared to simple aggregation.

Overconfidence Bias

Overconfidence refers to the tendency to overestimate one's predicted ability relative to performance. Studies show that people often "overshoot" when assigning probabilities to the chance that their answers are correct (Lichtenstein et al., 1977). While confidence does loosely track with performance, people still trend toward overestimation of ability (Koriat et al., 1980). Our analysis below suggests that neither finding holds for this particular dataset, as people consistently underestimate their performance, and underestimators outperform over estimators.

For this analysis, we took the cleaned dataframe from the "wisdom of crowds" assessment and further removed users who did not enter an estimation of their eventual performance. This dropped the dataset from 64,427 to 49,762 subjects. We then pivoted the data frame from a "wide" format (with a separate column for each of the 20 questions), to a "long" format (where every user had 20 rows, one for each question). This affords simpler analysis, as we can easily group by question and subject.

Despite expectations, we find that people do not overestimate their performance within this dataset. Only 29% of users overestimate their eventual score, whereas 62% underestimate themselves and 9% accurately perceive their performance. The distribution of errors are as follows, where overconfidence is measured by subtracting one's actual score from an estimate. (A negative score therefore implies underestimation.) See Figure 5.

Neither men nor women overestimate themselves on average in this task, although men are slightly more confident. The former has an average score of -1.51 and the latter has an average of -2.19. The distributions are otherwise similar. See Figure 6.

There appears to be a loose nonlinear trend in how age predicts overconfidence. Counterintuitively, subjects become less confident until age 18, and then they gradually start to develop better assessments of their ability. (In other words, their overconfidence approaches 0). After 50, accurate self-assessment dips once more, although in the direction of underestimation. See Figure 7.

People who report at least one year in a career that involved "reading people's expressions" underestimate themselves (with an average score of -1.73), while those without such careers underestimate themselves by a greater extent (with an average score of -2.15). This might suggest that experience leads to more accurate self-assessment. See Figure 8.

On average, overconfident subjects performed worse, with an average of 13.81 correct answers. Under estimators were the best performers, with an average of 16.07 correct answers, and accurate self-assessors on average correctly answered 15.02 prompts. See Figure 9.

Overall, we found that within this sample we saw general underestimation of ability across genders, careers, and ages. Men were less prone to underestimation than women, professionals were less prone to underestimation than non-professionals, and confidence in ability grows roughly from the ages 18 to 55. Additionally, the under-estimators wound up outperforming both the accurate and over-estimators.

References

Adams, J. K., & Adams, P. A. Realism of confidence judgments. Psychological Review, 1961, 68, 33-45.

Ashton, A. H., & Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. Management Science, 31, 1499-1508.

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. OR, 451-468.

Devore, J. L. (2016). Probability and statistics for engineering and the sciences (Ninth edition). Boston, MA: Cengage Learning.

Galton, F. (1907). Vox populi. Nature, 75(1949), 450-451.

Galton, F. (1907b). The Ballot-box. Nature, 75(1952), 509.

Kajdasz, J. E. (2014). A demonstration of the benefits of aggregation in an analytic task. International Journal of Intelligence and Counterintelligence, 27(4), 752-763.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118. doi: 10.1037/0278-7393.6.2.107

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art. In H. Jungermann & G. deZeeuw (Eds.), Decision making and change in human affairs.

Rosenberg, L. (2017). Human Swarms amplify accuracy in Honesty Detection. Presented at the Collective Intelligence Conference, Brooklyn, NY. Retrieved from http://unanimous.ai/wp-content/uploads/Human-Swarms-and-Honesty-Detection-CI-2017-Rosenberg.pdf

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. Journal of Behavioral Decision Making, 18, 29-53.

Spann, M., & Skiera, B. (2003). Internet-based virtual stock markets for business forecasting. Management Science, 49(10), 1310-1326.

Surowiecki, J. (2005). The Wisdom of Crowds. New York: Anchor Books.

Figures

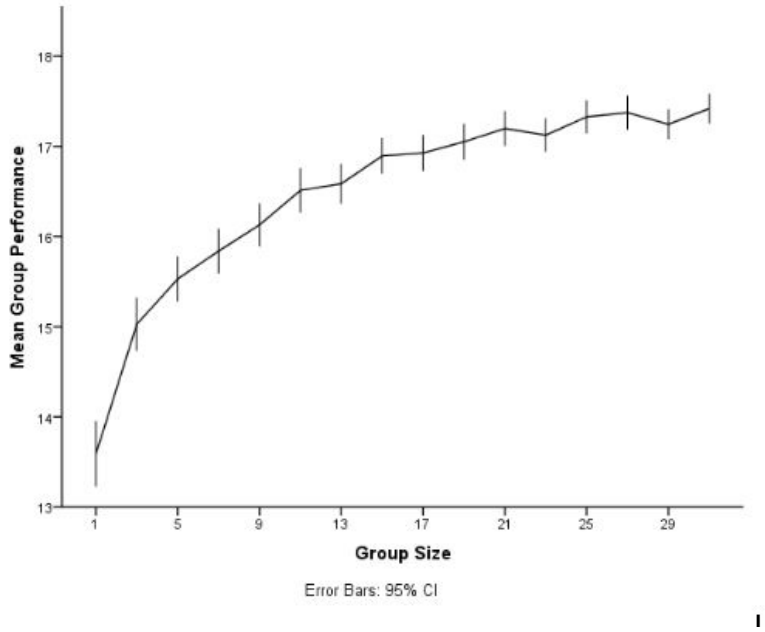

Figure 1: Group Performance for groups of various sizes on Smile Video Task. From Kajdasz (2014)

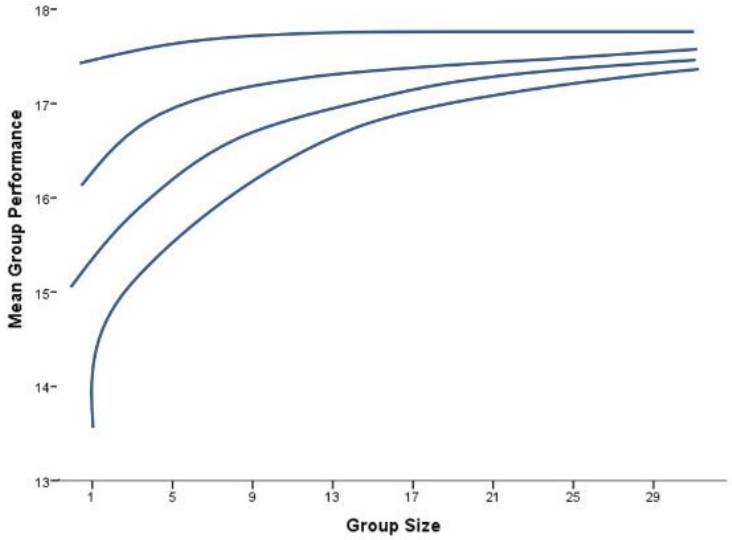

Figure 2: Hypothesized Performance for Groups of Various Sizes on Tasks of Varying Difficulty

Individual Mean Accuracy Rate for Each Smile Video (n = 64,427)

Video	1	2	3	4	5	6	7	8	9	10
Accuracy (%)	61.60	88.88	58.22	57.08	74.30	78.62	79.36	81.12	76.23	76.80

Video	11	12	13	14	15	16	17	18	19	20	*GM*
Accuracy (%)	72.54	75.24	81.23	71.68	78.00	77.23	55.82	93.81	85.41	63.47	74.33

Figure 3: Mean individual Accuracy for each smile video

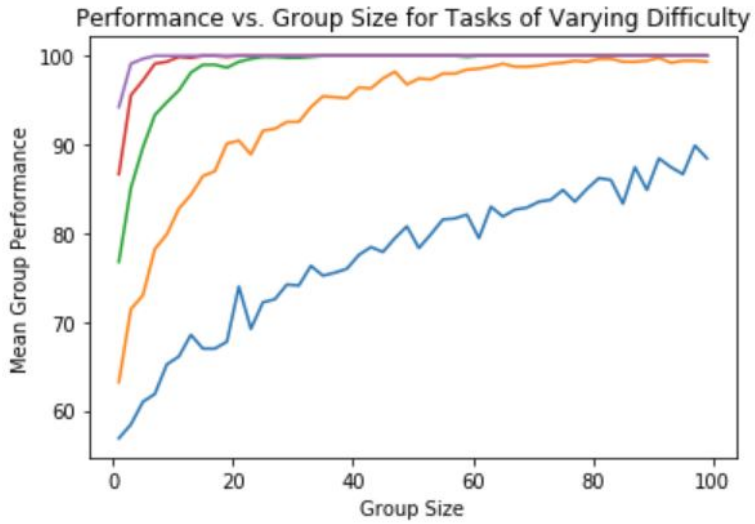

Figure 4: Observed Performance for Nominal Groups of Various Sizes on Tasks of Varying Difficulty

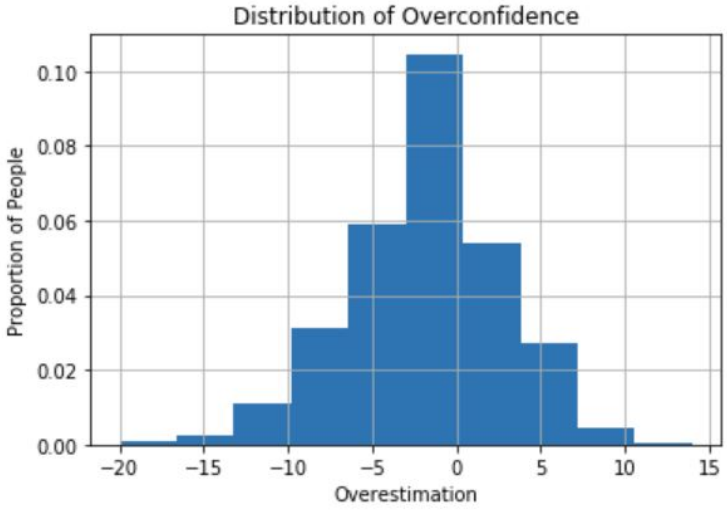

Figure 5: Overconfidence Exhibted in Participants (predicted accuracy - actual accuracy)

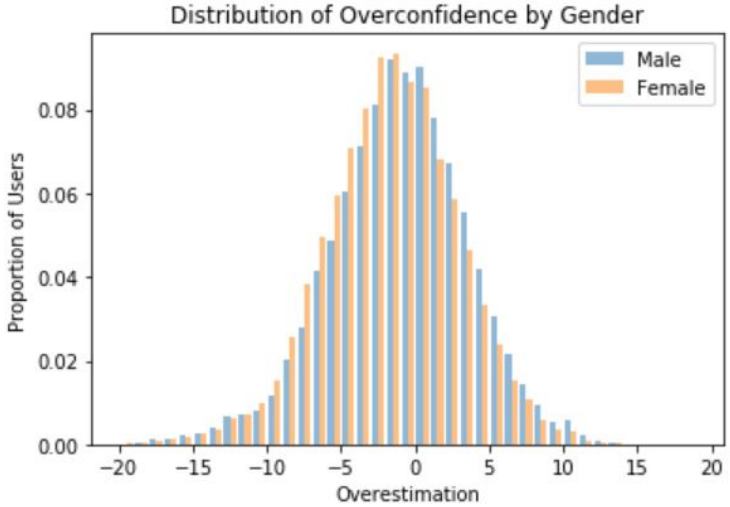

Figure 6: Overconfidence by Gender

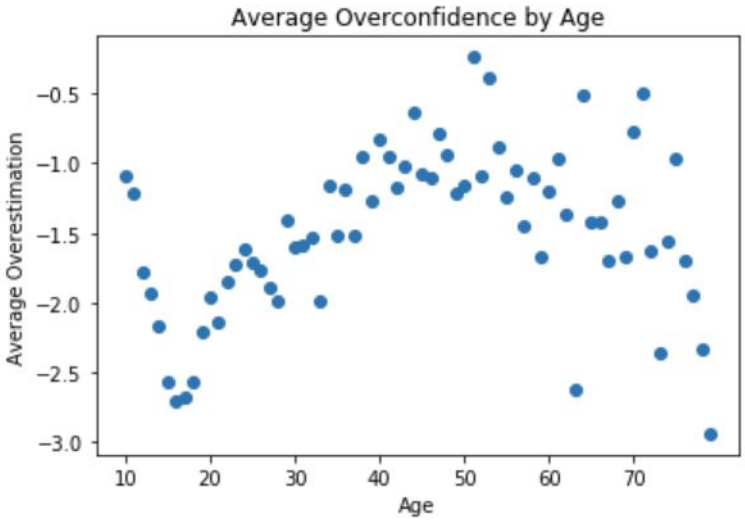

Figure 7: Overconfidence by Age

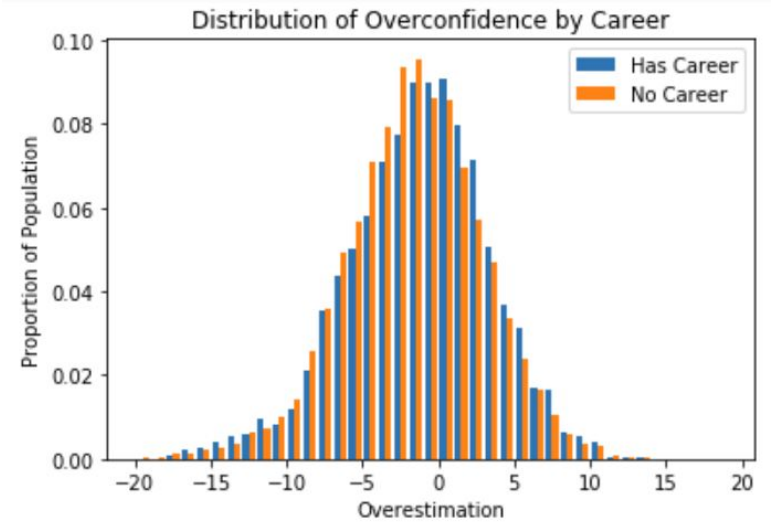

Figure 8: Overconfidence by Career

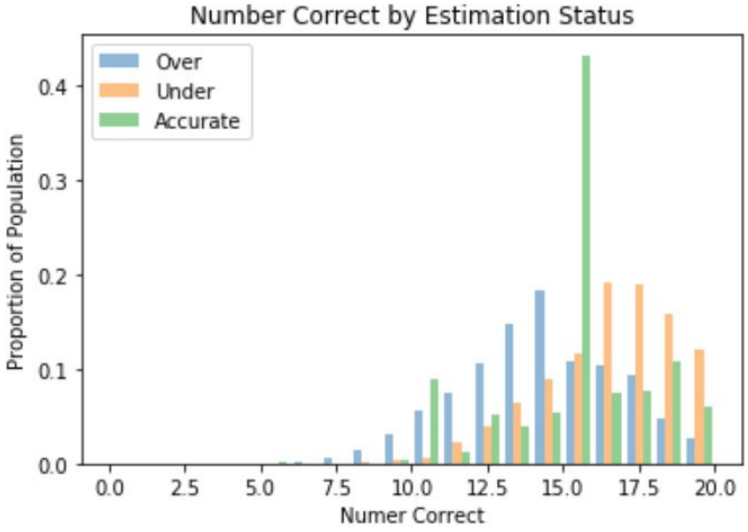

Figure 9: Perfromance observed groupd by Overconfidence