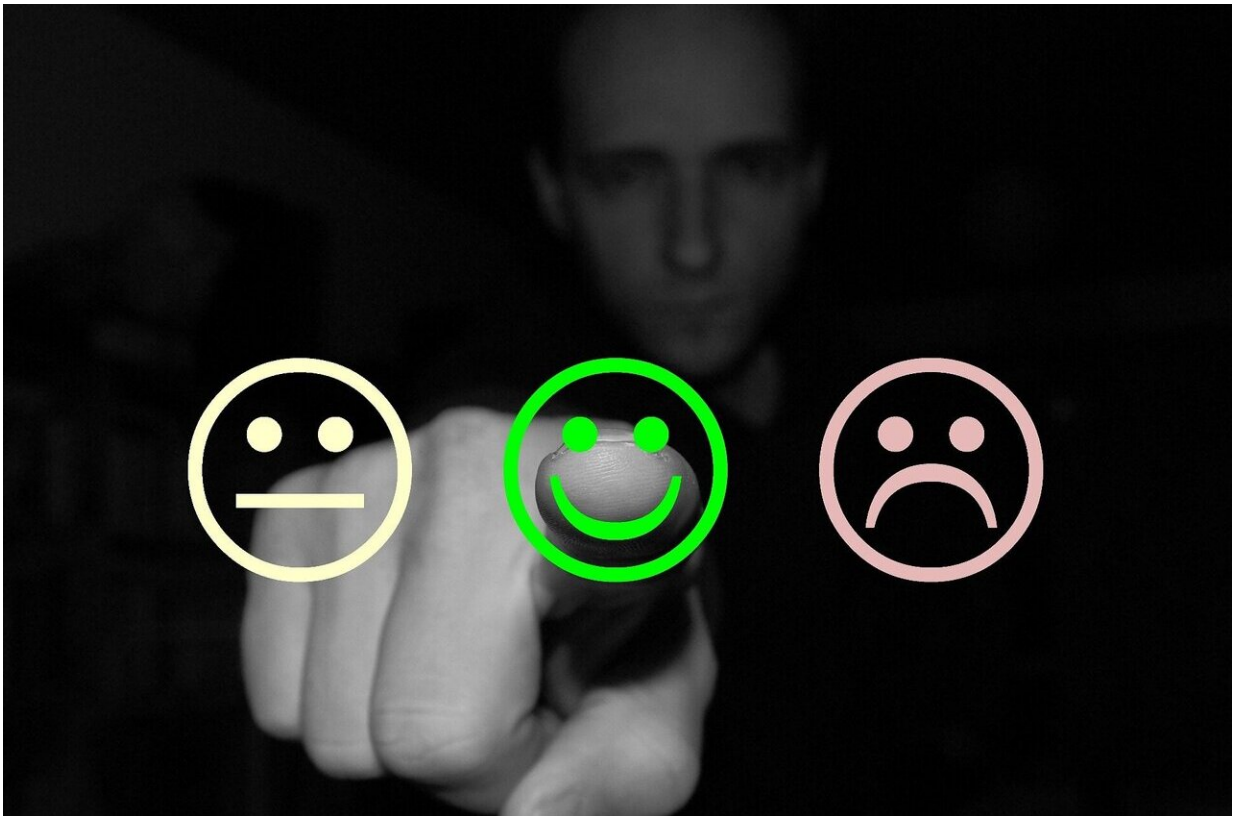# Yelp ratings get better when they cost something—like time

August 1 2019, by Fuller-Wright

Credit: CC0 Public Domain

An international team of researchers tested the idea that free online ratings are less trustworthy than those that have some cost to them, drawing from the ecological theory known as "costly signaling theory."

The theory suggests that if leaving a review carries some price—whether money or time or energy—it will result in more accurate ratings. In ecology, costly signaling theory argues that displays that "cost" more—like elaborate peacock tails, or strenuous displays of hunger from baby birds—are more likely to reflect reality. A colorful tail denotes a healthy peacock, and a chick with a full belly won't waste the energy to shout for more food.

But Princeton sociologist Dalton Conley and his colleagues are the first to apply this theory to Yelp or Uber and their ratings systems. By testing a series of weighted ratings tools in the context of a video game, they found that low-effort ratings were less accurate than those that cost a few extra seconds to use. They concluded that e-commerce sites should redesign their interfaces to impose time costs on raters of products or services.

"Simply put: making rating goods or services as easy as possible, as many e-commerce sites try to do, is counterproductive," said Conley, Princeton's Henry Putnam University Professor in Sociology and a faculty affiliate at the Office of Population Research and the Center for Health and Wellbeing, who is the senior author on a recent paper in the Proceedings of the National Academies of Science. "Ditto for forcing everyone to give a rating. Ratings are more accurate instead when they cost something to give."

He continued: "The intuition of Uber and other e-commerce sites is likely wrong. There's a reason that the peacock's feathers are so costly to produce: their cost assures an honest signal of reproductive fitness."

Or, as co-author Lucas Parra put it: "Online ratings are worthless, aren't they? Unless they incur some cost on the raters!" Parra is the Harold Shames Professor of Biomedical Engineering at the City College of New York.

Conley, Parra and their team of co-authors argued that even if there is little motivation to cheat with online ratings—there's no obvious incentive to leave a one-star review of a place we liked, or a five-star review of a dump—there is, at best, little direct benefit to raters who provide accurate assessments, suggesting that people are likely to provide low-quality information.

They decided to test the theory by imposing a "cost" to providing information—and higher costs on extreme ratings—to see if they could eliminate or reduce the number of dishonest, average-skewing one-star and five-star ratings.

So they created some video games, and recruited players from Amazon's Mechanical Turk.

In one typical game, players maneuvered a car to collect coins, knowing that they would receive one cent of real-life payment for each digital coin collected. Roads were separated by lakes that could only be traversed with ferries. The first two ferry rides were used as a training set, with delays of 20 seconds and then 4 seconds, to set a common baseline for ferry performance evaluations. After that, the game randomly varied the delays and speeds of ferry services. The fastest ferries arrived immediately and crossed the lake within 2 seconds, while the slowest ferries were both delayed in arrival and slow-moving, requiring a total of 40 seconds to cross a lake.

At the end of each ferry ride, players had to rate the ferry service on a scale of 0 to 100 before they could move on. Those ratings became the data for the research team. The in-game ratings tool used a weighted slide bar with digital "friction" for every point that a player moved away from a previously determined average rating. In other words, the more extreme your score, the more seconds you spent pushing the bar up or down.

Total gameplay was limited to 15 minutes, so players were motivated to submit their rankings as quickly as possible so they could go back to collecting their monetary rewards. Players rode an average of 17 ferries per game, allowing the researchers to measure correlations between their subjective ratings and the ferries' objective service (measured as total time to take the ferry), both within and across subjects.

They found that their weighted slide bar led to more reliable crowd estimates of quality than an unweighted click bar, where all scores from 0 to 100 could be given by an instant click on the screen—where all ratings were equally "cheap."

Their results have implications for the ubiquitous requests for ratings within e-commerce, and their approach can be generalized and tested in a variety of large-scale online communication systems, said the researchers.

The team hadn't set out to test ratings, said Conley. They were originally interested in online learning, "but in the course of experiments we realized that the ratings data we were getting … were not very good, so we set out to improve that problem."

They were surprised to find that reducing the cost of ratings actually backfired. Classic economic theory suggests that minimizing cost would yield the best results, but their data shows the opposite.

In short, Uber needs to slow down its rating tool, Conley said. "Converting the rating device from a simple click to a slider, where giving very high or low scores gets difficult due to the slider slowing down as the user gets farther out in either direction, yields better scoring distributions. Only highly motivated raters will provide extreme scores."

More information: Ofer Tchernichovski et al. Crowd wisdom

enhanced by costly signaling in a virtual rating system, *Proceedings of the National Academy of Sciences* (2019). DOI: 10.1073/pnas.1817392116