PERSONAL TECHNOLOGY

When 4.3 Stars Is Average: The Internet's Grade-Inflation Problem

Netflix is going with simpler thumbs-up or thumbs-down reviews while online star ratings for many products have lost their meaning

By

GEOFFREY A. FOWLER
April 5, 2017 1:43 p.m. ET

You know what I'd award zero stars? Most of the star ratings you see online.

You've got to dole out stars to your Uber driver, the movie you streamed, that Ben & Jerry's you ordered—and the dude who delivered it.

And what do you get back? I ran a little experiment tallying all the duct tape sold on Amazon. Of the 250 types and sizes, the average rating was 4.2 stars. On Yelp , I looked up gelato joints in San Francisco: More than half get either 4.5 or 5 stars. And I can't recall ever seeing an Uber driver with a rating lower than 4.3 stars.

— ADVERTISEMENT —

But wait, there's less: Online product ratings average about 4.3 stars all together, says PowerReviews, which runs ratings for more than 1,000 online shops.

MORE FROM TECH

Smart Lights Are a Bright Idea, With Alexa and Siri

Reliable voice control makes smart lighting easier than ever. Geoffrey A. Fowler explains the right—and wrong—ways to improve your home lighting with connected bulbs, plugs and switches.

CLICK TO READ STORY

Netflix is going with a simple thumbs-up or thumbs-down ratings system for reviews of its movies and other content. Netflix says it decided to replace the old five-star system because asking users to make fewer decisions results in more ratings.
PHOTO: NETFLIX

No doubt, online ratings can help weed out bad products, put sketchy restaurants out of business and keep dangerous drivers off the road. Reviews are a pillar of online culture, and there are still ways to extract useful info from them.

But I refuse to accept that everything on the internet is above average.

On Wednesday, Netflix NFLX -1.29% ▼ officially put its five-star system out to pasture. After extensive testing, it's replacing it with a simple thumbs up and thumbs down. "You get more ratings when you have fewer decision points," says Todd Yellin, Netflix's vice president of product innovation. And more data, combined with actual viewing behavior, allows Netflix to make personalized suggestions, represented as a percent match. It's like a dating site for movies.

Thumbs and recommendations aren't going to work for every app. But neither should star ratings, borrowed from professional critics and now often misused as a means to harvest the wisdom of the crowd.

People need more help interpreting the results. "Is three stars good or bad? I can't think of a platform that has nailed it," says Michael Luca, an assistant professor at Harvard Business School who studies information design.

Blame Yourself

It's partly our fault: We're far too nice.

The internet is known for turning people mean, but evidence shows we're much more likely to give positive ratings than bad ones. Yelp says 46% of the reviews we give local businesses are five stars. That may just be human nature: You chose this restaurant, so how could it be anything less than a five-star decision? Only every once in a while do you have a terrible experience you want to warn others about with a one-star rating.

Fistful of Fives

On Yelp, reviewers tend to be kind to businesses, giving out 5 stars most often. The rarest rating? 2 stars.

Distribution of Yelp's star ratings

★★★★★   **46%**

★★★★   **22%**

★★★   **10%**

★★   **7%**

★   **15%**
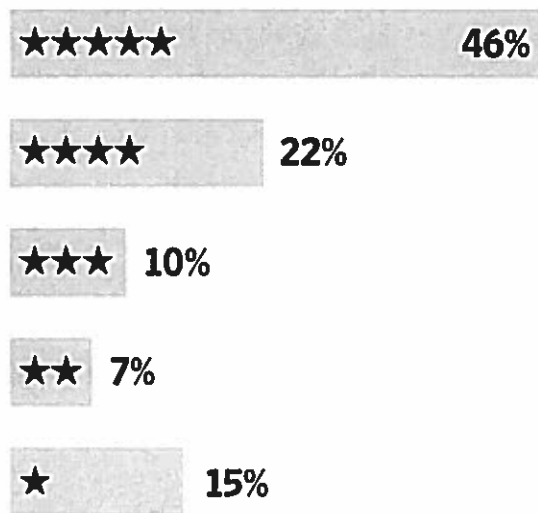
Source: the company

Many of us even lie to ourselves when we put on the "critic" hat, Mr. Yellin told me. Raise your hand if you're guilty of giving an Oscar-nominated film five stars because it was "important"... even if it actually put you to sleep.

It gets worse when apps use ratings to evaluate workers. Uber drivers can get the boot for relatively minor ratings dips, though the company won't specify the cutoff. So it feels socially awkward to give less than five stars, even if your driver's car kinda smells. Uber's stars aren't for finding good drivers, they're for flagging bad ones.

Yelp is using the same stars to do something very different: surface the best local businesses. But when every gelato shop has an above-par rating, we need other factors to pick a winner. That's one reason Yelp requires participants to leave written reviews that get closer to the heart of the matter. The downside: We have to trust the taste of people who have time to write Yelp reviews.

And then there's cheating. Apps inflate ratings by first asking if users are happy, then prodding them to write a review. (Not happy? Here's customer service.) I've seen Uber drivers offer to end the ride early—shaving a bit off the price—in exchange for a five-star rating. And that's not even counting the reviews by paid shills or internet trolls.

Better Ways

The good news is, many fixes are under way.

Amazon brought lawsuits against over 1,000 defendants for abuses such as purchasing fake reviews, and last year it changed its star system to give more weight to the most helpful reviews and those from verified purchasers. Airbnb discovered it got much more accurate reviews when it made them double-blind in 2014: Hosts and guests don't get to see each others' reviews before posting.

There's a tension between simplifying ratings to get more people involved, and asking for more details to make ratings more useful. When Netflix tested thumbs as an alternative to stars, it found participation rates doubled to 40%. Despite the loss of star-specific granularity, Netflix says people are also more honest with thumbs, which allows its software to make suggestions that actually stick.
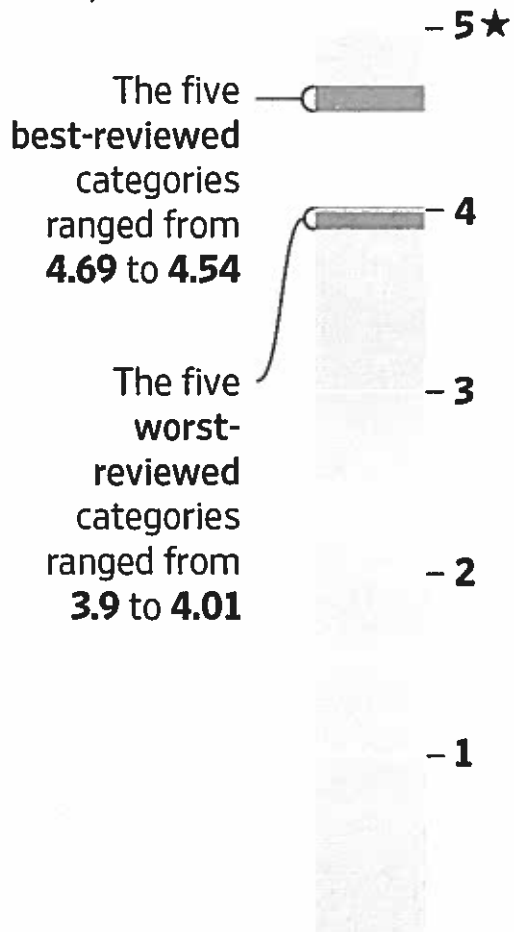
Everything Is Awesome

The average ratings of the most positively and negatively reviewed products are not far apart, says PowerReviews, which runs ratings systems for over 1,000 retailers.

— 5★

The five best-reviewed categories ranged from **4.69** to **4.54**
— 4

— 3

The five worst-reviewed categories ranged from **3.9** to **4.01**
— 2

— 1

Top categories are digital SLR cameras (4.69), camera lenses (4.6), electric guitars (4.6), soaps, sanitizers, lotions, wipes (4.54) and dog food (4.54). Bottom categories are bras (3.9), cigars (3.92), mascara (3.96), cell phones (3.96) and swimwear (4.01)

Source: PowerReviews

THE WALL STREET JOURNAL.

Uber also tested thumbs, along with smiley-face emojis, but says the alternatives made riders even more overly positive—and resulted in lost feedback for drivers. Uber did add menus to its five-star system for rider compliments and specific criticisms.

To make star ratings less inscrutable, TripAdvisor presents both averages and rankings. San Francisco's Maritime Museum has 4.5 stars, for example—but is actually only the city's 152nd best attraction.

Yelp's search results are ordered not just by average rating, but also by the number of reviews, how recent they are and your distance from the business. Its top-listed SF taco joint, El Farolito, has just four stars... but more than 4,000 reviews.

Replacing star systems won't be easy, because we have become accustomed to them. "All kinds of companies have attempted to create other scales, and most of them end up reverting back to five-star ratings," says Matt Moog, chief executive of PowerReviews.

What we really want is for these companies to get us. Alas, ratings get better when companies track more about what you are doing—whether you actually bought the product or ate at the restaurant. Netflix's insights are brilliant because the company knows what you stopped watching after 10 minutes.

A personalized recommendation is better than a four-point-whatever star rating any day of the week, but you only get it when you're really plugged in.

With online reviews more influential than ever, you need to read carefully to extract the truth:

- **Be wary of 5 stars.** Perfect scores mean not enough reviews. We're most likely to buy in the 4.2- to 4.5- star range, PowerReviews says.

- **Don't assume 3 is average.** Norms will vary by category. Eyeball the range of ratings, then grade on a curve.

- **Check the number of reviews.** How popular a product is and how recent its last review are strong indicators of quality.

- **Read written reviews.** One- and two-star explanations in particular can help you choose.

- **Log in.** When apps know who you are, they can make ratings more personalized.

MORE FROM POLITICS