Designing Better Online Review Systems

How to create ratings that buyers and sellers can trust
by Geoff Donaker, Hyunjin Kim, and Michael Luca
Designing Better Online Review Systems

How to create ratings that buyers and sellers can trust

AUTHORS

Geoff Donaker
Former COO, Yelp

Hyunjin Kim
Doctoral student, Harvard Business School

Michael Luca
Associate professor, Harvard Business School

ILLUSTRATOR SEAN MCCABE
Online reviews are transforming the way consumers choose products and services:

We turn to TripAdvisor to plan a vacation, Zocdoc to find a doctor, and Yelp to find new restaurants. Review systems play a central role in online marketplaces such as Amazon and Airbnb as well. More broadly, a growing number of organizations—ranging from Stanford Health Care to nine of the 10 biggest U.S. retailers—now maintain review ecosystems to help customers learn about their offerings.

Managed well, a review system creates value for buyers and sellers alike. Trustworthy systems can give consumers the confidence they need to buy a relatively unknown product, whether a new book or dinner at a local restaurant. For example, research by one of us (Mike) found that higher Yelp ratings lead to higher sales. This effect is greater for independent businesses, whose reputations are less well established.

Reviews also create a feedback loop that provides suppliers with valuable information: For example, ratings allow Uber to remove poorly performing drivers from its service, and they can give producers of consumer goods guidance for improving their offerings.

But for every thriving review system, many others are barren, attracting neither reviewers nor other users. And some amass many reviews but fail to build consumers’ trust in their informativeness. If reviews on a platform are all positive, for example, people may assume that the items being rated are all of high quality—or they may conclude that the system can’t help them differentiate the good from the bad. Reviews can be misleading if they provide an incomplete snapshot of experiences. Fraudulent or self-serving reviews can hamper platforms’ efforts to build trust. Research by Mike and Georgios Zervas has found that businesses are especially likely to engage in review fraud when their reputation is struggling or competition is particularly intense.

Behind many review-system failures lies a common assumption: that building these systems represents a technological challenge rather than a managerial one. Business leaders often invest heavily in the technology behind a system but fail to actively manage the content, leading to common...
problems. The implications of poor design choices can be severe: It’s hard to imagine that travelers would trust Airbnb without a way for hosts to establish a reputation (which leans heavily on reviews), or that shoppers could navigate Amazon as seamlessly without reviews. As academics, Hyunjin and Mike have researched the design choices that lead some online platforms to succeed while others fail and have worked with Yelp and other companies to help them on this front (Hyunjin is also an economics research intern at Yelp). And as the COO of Yelp for more than a decade, Geoff helped its review ecosystem become one of the world’s dominant sources of information about local services.

In recent years a growing body of research has explored the design choices that can lead to more-robust review and reputation systems. Drawing on our research, teaching, and work with companies, this article explores frameworks for managing a review ecosystem—shedding light on the issues that can arise and the incentives and design choices that can help to avoid common pitfalls. We’ll look at each of these issues in more detail and describe how to address them.

Not Enough Reviews

When Yelp began, it was by definition a new platform—a ghost town, with few reviewers or readers. Many review systems experience a shortage of reviews, especially when they’re starting out. While most people read reviews to inform a purchase, only a small fraction write reviews on any platform they use. This situation is exacerbated by the fact that review platforms have strong network effects: It is particularly difficult to attract review writers in a world with few readers, and difficult to attract readers in a world with few reviews.

We suggest three approaches that can help generate an adequate number of reviews: seeding the system, offering incentives, and pooling related products to display their reviews together. The right mix of approaches depends on factors such as where the system is on its growth trajectory, how many individual products will be included, and what the goals are for the system itself.

Seeding reviews. Early-stage platforms can consider hiring reviewers or drawing in reviews from other platforms (through a partnership and with proper attribution). To create enough value for users in a new city to start visiting Yelp and contributing their own reviews, the company recruited paid teams of part-time “scouts” who would add personal photos and reviews for a few months until the platform caught on. For other businesses, partnering with platforms that specialize in reviews can also be valuable—both for those that want to create their own review ecosystem and for those that want to show reviews but don’t intend to create their own platform. Companies such as Amazon and Microsoft pull in reviews from Yelp and other platforms to populate their sites.

For platforms looking to grow their own review ecosystem, seeding reviews can be particularly useful in the early stages because it doesn’t require an established brand to incentivize activity. However, a large number of products or services can make it costly, and the reviews that you get may differ from organically generated content, so some platforms—depending on their goals—may benefit from swiftly moving beyond seeding.

Offering incentives. Motivating your platform’s users to contribute reviews and ratings can be a scalable option and can also create a sense of community. The incentive you use might be financial: In 2014 Airbnb offered a $25 coupon in exchange for reviews and saw a 6.4% increase in review rates. However, nonfinancial incentives—such as in-kind gifts or status symbols—may also motivate reviewers, especially if your brand is well established. In Google’s Local Guides program, users earn points any time they contribute something to the platform—writing a review, adding a photo, correcting content, or answering a question. They can convert those points into rewards ranging from early access to new Google products to a free 1TB upgrade of Google Drive storage. Yelp’s “elite squad” of prolific, high-quality reviewers receive a special designation on the platform along with invitations to private parties and events, among other perks.

Financial incentives can become a challenge if you have a large product array. But a bigger concern may be that if they aren’t designed well, both financial and nonfinancial incentives can backfire by inducing users to populate the system with fast but sloppy reviews that don’t help other customers.
Pooling products. By reconsidering the unit of review, you can make a single comment apply to multiple products. On Yelp, for example, hairstylists who share salon space are reviewed together under a single salon listing. This aggregation greatly increases the number of reviews Yelp can amass for a given business, because a review of any single stylist appears on the business’s page. Furthermore, since many salons experience regular churn among their stylists, the salon’s reputation is at least as important to the potential customer as that of the stylist. Similarly, review platforms may be able to generate more-useful reviews by asking users to review sellers (as on eBay) rather than separating out every product sold.

Deciding from the outset whether and how to pool products in a review system can be helpful, because it establishes what the platform is all about. (Is this a place to learn about stylists or about salons?) Pooling becomes particularly attractive as your product space broadens, because you have more items to pool in useful ways.

A risk to this approach, however, is that pooling products to achieve more reviews may fail to give your customers the information they need about any particular offering. Consider, for example, whether the experience of visiting each stylist in the salon is quite different and whether a review of one stylist would be relevant to potential customers of another.

Amazon’s pooling of reviews in its bookstore takes into account the format of the book a reader wants to buy. Reviews of the text editions of the same title (hardback, paperback, and Kindle) appear together, but the audiobook is reviewed separately, under the Audible brand. For customers who want to learn about the content of the books, pooling reviews for all audio and physical books would be beneficial. But because audio production quality and information about the narrator are significant factors for audiobook buyers, there may be a benefit to keeping those reviews separate.

All these strategies can help overcome a review shortage, allowing content development to become more self-sustaining as more readers benefit from and engage with the platform. However, platforms have to consider not only the volume of reviews but also their informativeness—which can be affected by selection bias and gaming of the system.

Selection Bias

Have you ever written an online review? If so, what made you decide to comment on that particular occasion? Research has shown that users’ decisions to leave a review often depend on the quality of their experience. On some sites, customers may be likelier to leave reviews if their experience was good; on others, only if it was very good or very bad. In either case the resulting ratings can suffer from selection bias: They might not accurately represent the full range of customers’ experiences of the product. If only satisfied people leave reviews, for example, ratings will be artificially inflated. Selection bias can become even more pronounced when businesses nudge only happy customers to leave a review.

EBay encountered the challenge of selection bias in 2011, when it noticed that its sellers’ scores were suspiciously high: Most sellers on the site had over 99% positive ratings. The company worked with the economists Chris Nosko and Steven Tadelis and found that users were much likelier to leave a review after a good experience: Of some 44 million transactions that had been completed on the site, only 0.39% had negative reviews or ratings, but more than twice as many (1%) had an actual “dispute ticket,” and more than seven times as many (3%) had prompted buyers to exchange messages with sellers that implied a bad experience. Whether or not buyers decided to review a seller was in fact a better predictor of future complaints—and thus a better proxy for quality—than that seller’s rating.

EBay hypothesized that it could improve buyers’ experiences and thus sales by correcting for raters’ selection bias and more clearly differentiating higher-quality sellers. It reformulated seller scores as the percentage of all of a seller’s transactions that generated positive ratings (instead of the percentage of positive ratings). This new measure yielded a median of 67% with substantial spread in the distribution of scores—and potential customers who were exposed to the new scores were more likely than a control group to return and make another purchase on the site.

By plotting the scores on your platform in a similar way, you can investigate whether your ratings are skewed, how severe the problem may be, and whether additional data might help you fix it. Any review system can be crafted to

On some sites, customers may be likelier to leave reviews if their experience was good; on others, only if it was very good or very bad.
mitigate the bias it is most likely to face. The entire review process—from the initial ask to the messages users get as they type their reviews—provides opportunities to nudge users to behave in less-biased ways. Experimenting with design choices can help show how to reduce the bias in reviewers’ self-selection as well as any tendency users have to rate in a particular way.

**Require reviews.** A more heavy-handed approach requires users to review a purchase before making another one. But tread carefully: This may drive some customers off the platform and can lead to a flood of noninformative ratings that customers use as a default—creating noise and a different kind of error in your reviews. For this reason, platforms often look for other ways to minimize selection bias.

**Allow private comments.** The economists John Horton and Joseph Golden found that on the freelancer review site Upwork, employers were reluctant to leave public reviews after a negative experience with a freelancer but were open to leaving feedback that only Upwork could see. (Employers who reported bad experiences privately still gave the highest possible public feedback nearly 20% of the time.) This provided Upwork with important information—about when users were or weren’t willing to leave a review, and about problematic freelancers—that it could use either to change the algorithm that suggested freelancer matches or to provide aggregate feedback about freelancers. Aggregate feedback shifted hiring decisions, indicating that it was relevant additional information.

**Design prompts carefully.** More generally, the reviews people leave depend on how and when they are asked to leave them. Platforms can minimize bias in reviews by thoughtfully designing different aspects of the environment in which users decide whether to review. This approach, often referred to as choice architecture—a term coined by Cass Sunstein and Richard Thaler (the authors of *Nudge: Improving Decisions About Health, Wealth, and Happiness*)—applies to everything from how prompts are worded to how many options a user is given.

In one experiment we ran on Yelp, we varied the messages prompting users to leave a review. Some users saw the generic message “Next review awaits,” while others were asked to help local businesses get discovered or other consumers to find local businesses. We found that the latter group tended to write longer reviews.

**Fraudulent and Strategic Reviews**

Sellers sometimes try (unethically) to boost their ratings by leaving positive reviews for themselves or negative ones for their competitors while pretending that the reviews were left by real customers. This is known as astroturfing. The more influential the platform, the more people will try to astroturf.

Because of the harm to consumers that astroturfing can do, policymakers and regulators have gotten involved. In 2013 Eric Schneiderman, then the New York State attorney general, engaged in an operation to address it—citing our research as part of the motivation. Schneiderman’s office announced an agreement with 19 companies that had helped write fake reviews on online platforms, requiring them to stop the practice and to pay a hefty fine for charges including false advertising and deceptive business practices. But, as with shoplifting, businesses cannot simply rely on law enforcement; to avoid the pitfalls of fake reviews, they must set up their own protections as well. As discussed in a paper that Mike wrote with Georgios Zervas, some companies, including Yelp, run sting operations to identify and address companies trying to leave fake reviews.

A related challenge arises when buyers and sellers rate each other and craft their reviews to elicit higher ratings from the other party. Consider the last time you stayed in an Airbnb. Afterward you were prompted to leave a review of the host, who was also asked to leave a review of you. Until 2014, if you left your review before the host did, he or she could read it before deciding what to write about you. The result? You might think twice before leaving a negative review.

Platform design choices and content moderation play an important role in reducing the number of fraudulent and strategic reviews.

**Set rules for reviewers.** Design choices begin with deciding who can review and whose reviews to highlight. For example, Amazon displays an icon when a review is from a verified purchaser of the product, which can help consumers screen for potentially fraudulent reviews. Expedia goes further and allows only guests who have booked through its platform to leave a review there. Research by Dina Mayzlin, Yaniv Dover, and Judith Chevalier shows that such a policy can reduce the number of fraudulent reviews. At the same time, stricter rules about who may leave a review can be a blunt instrument that significantly diminishes the number of genuine reviews and reviewers. The platform must decide whether the benefit of reducing potential fakes exceeds the cost of having fewer legitimate reviews.
Platforms also decide when reviews may be submitted and displayed. After realizing that nonreviewers had systematically worse experiences than reviewers, Airbnb implemented a “simultaneous reveal” rule to deter reciprocal reviews between guests and hosts and allow for more-complete feedback. The platform no longer displays ratings until both the guest and the host have provided them and sets a deadline after which the ability to review expires. After the company made this change, research by Andrey Fradkin, Elena Grewal, and David Holtz found that the average rating for both guests and hosts declined, while review rates increased—suggesting that reviewers were less afraid to leave feedback after a bad experience when they were shielded from retribution. **Call in the moderators.** No matter how good your system’s design choices are, you’re bound to run into problems. Spam can slip in. Bad actors can try to game the system. Reviews that were extremely relevant two years ago may become obsolete. And some reviews are just more useful than others. Reviews from nonpurchasers can be ruled out, for example, but even some of those that remain may be misleading or less informative. Moderation can eliminate misleading reviews on the basis of their content, not just because of who wrote them or when they were written.

Content moderation comes in three flavors: employee, community, and algorithm. Employee moderators (often called community managers) can spend their days actively using the service, interacting online with other users, removing inappropriate content, and providing feedback to management. This option is the most costly, but it can help you quickly understand what’s working and what’s not and ensure that someone is managing what appears on the site at all times.

Community moderation allows all users to help spot and flag poor content, from artificially inflated reviews to spam and other kinds of abuse. Yelp has a simple icon that users can post to submit concerns about a review that harasses another reviewer or appears to be about some other business. Amazon asks users whether each review is helpful or unhelpful and employs that data to choose which reviews are displayed first and to suppress particularly unhelpful ones. Often only a small fraction of users will flag the quality of content, however, so a critical mass of engaged users is needed to make community flagging systems work.

The third approach to moderating content relies on algorithms. Yelp’s recommendation software processes dozens of factors about each review daily and varies the reviews that are more prominently displayed as “recommended.” In 2014 the company said that fewer than 75% of written reviews were recommended at any given time. Amazon, Google, and TripAdvisor have implemented review-quality algorithms that remove offending content from their platforms. Algorithms can of course go beyond a binary classification and instead assess how much weight to place on each rating. Mike has written a paper with Daisy Dai, Ginger Jin, and Jungmin Lee that explores the rating aggregation problem, highlighting how assigning weights to each rating can help overcome challenges in the underlying review process.

**Putting It All Together**

The experiences of others have always been an important source of information about product quality. The American Academy of Family Physicians, for example, suggests that people turn to friends and family to learn about physicians and get recommendations. Review platforms have accelerated and systematized this process, making it easier to tap into the wisdom of the crowd. Online reviews have been useful to customers, platforms, and policymakers alike. We have used Yelp data, for example, to look at issues ranging from understanding how neighborhoods change during periods of gentrification to estimating the impact of minimum-wage hikes on business outcomes. But for reviews to be helpful—to consumers, to sellers, and to the broader public—the people managing review systems must think carefully about the design choices they make and how to most accurately reflect users’ experiences.

Geoff Donaker is a former chief operating officer and board member of Yelp. Hyunjin Kim is a doctoral student in the strategy unit at Harvard Business School. Michael Luca is the Lee J. Styerlenger III Associate Professor of Business Administration at Harvard Business School and a coauthor (with Max H. Bazerman) of The Power of Experiments: Decision Making in a Data-Driven World (forthcoming from MIT Press).