Article

AMERICAN MARKETING ASSOCIATION

Journal of Marketing Research
2021, Vol. 58(4) 742-761

DOI: 10.1177/00222437211020274
journals.sagepub.com/home/mrj
SAGE

Interacting User-Generated Content Technologies: How Questions and Answers Affect Consumer Reviews

Shrabastee Banerjee, Chrysanthos Dellarocas, and Georgios Zervas

Abstract

This article studies the question and answer (Q&A) technology of electronic commerce platforms, an increasingly common form of user-generated content that allows consumers to publicly ask product-specific questions and receive responses, either from the platform or from other customers. Using data from a major online retailer, the authors show that Q&As complement consumer reviews: unlike reviews, questions are primarily asked prepurchase and focus on clarification of product attributes rather than discussion of quality; answers convey fit-specific information in a predominantly sentiment-free way. Drawing on these observations, the authors hypothesize that Q&As mitigate product fit uncertainty, leading to better matches between products and consumers and, therefore, improved product ratings. Indeed, when products suffering from fit mismatch start receiving Q&As, their subsequent ratings improve by approximately .1 to .5 stars, and the fraction of negative reviews that discuss fit-related issues declines. The extent of the rating increase due to Q&As is proportional to the probability that purchasers will experience fit mismatch without Q&A. These findings suggest that, by resolving product fit uncertainty in an e-commerce setting, the addition of Q&As can be a viable way for retailers to improve ratings of products that have incurred low ratings due to customer–product fit mismatch.

Keywords

e-commerce, Q&A, reputation systems, user-generated content

Online supplement: https://doi.org/10.1177/00222437211020274

Consumer reviews have been shown to influence purchase decisions in the context of both products and services and are widely adopted by brands and retailers (Chevalier and Mayzlin 2006; Luca 2016; Zhu and Zhang 2010). Recently, another form of user-generated content, questions and answers (Q&As), has been gaining traction with online retailers and review platforms. Q&A technology, which is typically implemented alongside reviews, enables consumers to ask specific questions about a product and receive answers from another consumer, the brand, or the platform itself. Q&A technology is now widely adopted by retailers and has been embraced by consumers.[1] Despite this increased usage, little is known about how this technology affects consumer decision making.

In this article, our aim is to fill this gap and examine the impact of Q&As on consumer decision making. Using data on consumer reviews and Q&As from a major U.K.-based electronic commerce platform spanning a period of five years, we show that Q&A technology resolves an important information problem and ultimately leads to better purchase decisions. The information problem arises from two sources of uncertainty that consumers face when trying to evaluate a product: quality uncertainty and fit uncertainty. Quality is a product-level characteristic that consumers agree on, whereas fit captures idiosyncratic preferences that are specific to individual consumers. Fit uncertainty is exacerbated in an online setting because consumers cannot interact physically with a product prior to purchase.

We hypothesize that products often receive low ratings not because of quality concerns but rather because of consumer–product mismatches owing to fit uncertainty, possibly resulting

[1] The Web Appendix shows examples of Q&As from various platforms. Amazon displays Q&As prominently above consumer reviews on each product page.

Shrabastee Banerjee is a doctoral candidate, Department of Marketing, Questrom School of Business, Boston University, USA (email: sban@bu.edu). Chrysanthos Dellarocas is Shipley Professor of Management, Questrom School of Business, Boston University USA (email: dell@bu.edu). Georgios Zervas is Associate Professor, Department of Marketing, Questrom School of Business, Boston University, USA (email: zg@bu.edu).

from inadequate or wrong information on a retailer's website, highly individualized fit requirements on behalf of consumers, or intrinsic product complexity.[2] To alleviate fit mismatch, online retailers have traditionally relied on consumer reviews. We posit that Q&A technology can act as an effective complement to reviews and can help resolve any residual fit uncertainty that reviews might fail to address. By comparing our Q&A and review corpora, we find substantive differences in both how consumers use these information sources and their contents. We find that, unlike reviews, Q&As primarily happen prepurchase, focus on clarification of product attributes rather than discussion of quality, and convey fit-specific information in a relatively concise and sentiment-free way. By contrast, because review text does not have a predefined structure that requires authors to comment on both quality and fit issues, it can be difficult for individual consumers to deduce fine aspects of product fit from this corpus. In these cases, Q&A technology can be a complement to consumer reviews because it enables individual consumers to inquire or read about their specific sources of uncertainty and receive answers before purchase. By addressing specific concerns about product features that may not come up in reviews,[3] Q&As can mitigate fit uncertainty before purchases happen. Thus, our main hypothesis is that Q&A technology can help resolve fit uncertainty where it exists, leading to greater consumer satisfaction postpurchase, which in turn results in higher product evaluations in the form of increased ratings.

A challenge that we face in testing our main hypothesis is identifying products that are more likely to suffer from fit mismatch. Because we do not directly observe whether a negative review is posted due to quality- or fit-related issues, we construct three different proxies for fit mismatch. Our first measure, motivated by the observation that mismatch causes low ratings, is the average rating of each product prior to the arrival of its first Q&A. While this measure is straightforward to compute, it is imperfect: some products might have low ratings due to quality issues rather than fit mismatch. Our second measure addresses this concern by taking into account the variance in ratings. High variance signals more heterogeneity in consumer preferences for certain attributes of the product, and therefore a higher likelihood of fit mismatch. Finally, our third measure takes into account the text content of negative reviews. We begin by asking human coders to read and categorize a sample of negative reviews into one of three categories: poor quality, poor fit, or other miscellaneous reasons (e.g., shipping concerns). We then use these human-labeled reviews to train a classifier that detects fit concerns. We apply this classifier to all negative reviews in our data to construct our third measure: the fraction of each product's reviews that discusses fit issues.

Using data from a major U.K. retailer, we estimate the effect of Q&As on subsequent product ratings by exploiting variation in the timing of Q&As posted for different products. We find that answering questions for products that suffer from fit mismatch increases their subsequent ratings by roughly .1 stars. Moreover, we find that the extent of this rating increase is proportional to the probability that purchasers experience fit mismatch for that product prior to Q&A. To provide evidence around our hypothesized mechanism—that Q&As lead to better matches between consumers and products—we estimate the impact of Q&As on the probability of products receiving negative reviews mentioning fit-related issues. We find that the fraction of negative reviews due to fit concerns declines following the arrival of Q&As, with the extent of this decline again being proportional to the probability of fit mismatch prior to Q&A.

To interpret these results causally, we need to assume that the timing of Q&As is not correlated with time-varying unobservables that can also affect product ratings. This assumption could be violated in our setting. In particular, unobserved marketing-related activities such as product page improvements, discounts, and promotions could attract more consumers to specific products, leading to Q&As. To the extent that these marketing activities are well-targeted, they could also lead to higher ratings. We approach these threats to validity in several ways. First, we use an auxiliary data set of browsing behavior to directly look for patterns suggestive of demand shocks. We find no changes in the volume of reviews or product page impressions around the time Q&As arrive. Next, we collect additional data from the Internet Archive (https://archive.org/), which allows us to look at historical snapshots of the product pages in our sample. From these data, we reestimate our main specifications controlling for historical prices, promotions, and product description length, and we find no change to our results. These robustness checks suggest that our results are not being driven by unobserved marketing-related activities. Finally, we check whether there is an influx of reviews whose contents address fit concerns coinciding with the arrival of Q&As, which would confound our attribution. We test for changes in the composition of review text around the time of the first answer and find no evidence that review contents change around the arrival of Q&As.

Overall, our findings suggest that, by resolving fit uncertainty in an e-commerce setting, the implementation of Q&A technology can be a viable way for retailers to improve product ratings, particularly for products that have suffered low ratings due to consumer–product fit mismatch.

Related Work

The impact of ratings and reviews on consumer behavior (most notably, purchase decisions) has been well-documented in the literature. For example, an online experiment shows that participants who consulted product recommendations selected these products twice as often as those who did not (Senecal and Nantel 2004). Online consumer ratings have also been found

[2] For some illustrative examples of reviews arising from poor quality versus poor fit, see the Web Appendix.
[3] For instance, "My studio flat door is 27 inches wide, would it come through the door?" or "Does this work with Nikon L820 Bridge Camera?"

to significantly influence product sales in the market for books (Chevalier and Mayzlin 2006). Similarly, in the domain of services, a one-star increase in a restaurant's Yelp rating led to 5%–9% increase in revenues (Luca 2016). Although in this article we do not directly look at purchases, these studies show that an increase in average ratings is a positive and managerially relevant outcome, because it widely correlates with downstream conversion.

Other studies have looked more deeply into the impact of different characteristics of reviews on sales. More helpful reviews and highlighted reviews have been found to have a stronger impact on sales (Dhanasobhon et al. 2007). Further, the impact of reviews on sales is stronger for less popular products and for customers who have greater internet experience (Zhu and Zhang 2010). The text content of reviews has also been established to be of importance above and beyond the corresponding numerical rating (Archak, Ghose, and Ipeirotis 2011).

In contrast to reviews, the role of user-generated Q&As in influencing conversion or related buyer behavior in an e-commerce setting has not been examined. Most of the work in the domain of Q&As has focused on question-answering communities, such as Quora and Stack Overflow. Questions examined in this area include how reputation relates to response volume, question difficulty, and answer quality on Stack Overflow (Lappas, Dellarocas, and Derakhshani 2017); how to model the satisfaction of information seekers in Q&A communities (Agichtein, Liu, and Bian 2009); what makes a "good" question in a community setting (Ravi et al. 2014); and so on. In terms of the interplay between Q&A-type communities and purchase behavior, research has shown that engagement in firm-operated online communities can lead customers to spend more on the firm (i.e., accrue more "social dollars"), with this effect being strongest for posters of community content and those with more social ties (Manchanda, Packard, and Pattabhiramaiah 2015). In such a setting, the source of economic benefit is viewed as social rather than informational. Our work highlights an alternative channel through which Q&A platforms might provide a benefit to firms if they are integrated within an e-commerce framework, namely by resolving fit mismatch and leading to higher consumer satisfaction. The most closely related work to our article that also examines the overlap between Q&As and reviews in an e-commerce setting develops an algorithm to show how existing reviews can be mined to answer questions on Amazon (McAuley and Yang 2016). However, the focus of this work is on developing and comparing the algorithm with other existing text-mining tools, and it does not investigate any causal questions that combine reviews and Q&As.

Our "Conceptual Framework" section highlights how consumers make use of reviews and Q&As when both are present simultaneously on the product page. Closely related to the constructs of horizontal and vertical differentiation (Tirole 1988), we posit that consumers are subject to two distinct varieties of uncertainty in an online setting: product quality uncertainty and product fit uncertainty. Broadly construed, product quality uncertainty is the consumer's difficulty in evaluating product quality and predicting how a product will perform in the future (Dimoka, Hong, and Pavlou 2012). Products may have inherent quality issues that are revealed only through prolonged product usage; thus, reviews can be a valuable avenue through which quality uncertainty is mitigated.

Product fit uncertainty, in contrast, arises because buyers cannot easily assess whether the product's characteristics match their requirements or tastes (Hong and Pavlou 2014; Kwark, Chen, and Raghunathan 2014). Fit uncertainty might thus lead to mismatched purchases and thereby attract low ratings even if the inherent product quality is good. While different consumers may have the same level of quality uncertainty with a certain amount of information, their level of product fit uncertainty may vary due to their particular needs and heterogeneous fit preferences. Thus, we posit that attribute-based Q&A content can alleviate fit uncertainty more directly and completely than review text alone.

Prior work has also explored various other avenues through which these uncertainties can potentially be addressed, but without reference to Q&A technology. For instance, using survey data from consumers, researchers observed that participation in online product forums reduces product fit uncertainty, whereas the use of online media on product pages reduces product quality uncertainty (Hong and Pavlou 2014). Our results relate to this work in the sense that we can think of Q&As as being similar to product forums that reduce fit uncertainty (in both cases, customers can bring up or read about specific concerns they have about a product). Fit uncertainty in an e-commerce setting can also be reduced with virtual reality widgets. For instance, in the context of apparel, it has been shown that offering virtual fitting rooms increases conversion, basket sizes, average price of purchased products, and revisits to the site while reducing fulfillment costs arising from returns and home try-on behavior (Gallino and Moreno 2018).

In the general domain of product returns, the two types of uncertainty have also been argued for: research has shown that product fit uncertainty is mitigated by offline inspection and visits to the store, whereas reviews can offer a strong quality signal, thereby mitigating quality uncertainty, both of which can reduce return rates (Sahoo, Dellarocas, and Srinivasan 2018). We posit that, apart from offline inspections and augmented reality apps (e.g., virtual trials), Q&As can be an effective tool with which retailers can reduce fit uncertainty.

In addition to using average ratings and review text to measure the probability of inherent fit mismatch for a product, we also make use of rating variance. It has been shown that niche products, which some consumers like but others dislike, can give rise to high variance (Sun 2012). We would thus expect Q&As to facilitate better-informed purchases for such products.

Finally, our setting differs from one in which quality and fit are more intrinsically linked and hard to disentangle (e.g., for books or movies). For instance, research has shown that reviews on Goodreads can influence the nature of product discovery and thus shape consumer choices by allowing consumers to find lesser-known products that match their taste, more so than simply identifying products of high quality. In such settings, the role of Q&As would be more nuanced, because there are fewer objective attributes, and open-ended reviews might be more helpful in terms of resolving fit uncertainty (Bondi 2019).

Conceptual Framework

Uncertainty in the context of e-commerce can be considered the result of two information problems: quality uncertainty and fit uncertainty. The industrial organization literature (Tirole 1988) defines quality as a product-level attribute that is commonly perceived by all consumers, whereas fit reflects aspects of utility that are specific to individual consumers and can be highly idiosyncratic. In modern electronic commerce platforms, a key mechanism for reducing quality uncertainty is product reviews contributed by past purchasers. Product reviews can also provide information about fit. Nevertheless, reviews are not as well-suited to reducing fit uncertainty because they typically mix discussion of several aspects of a consumer's experience with a product (e.g., performance, features, reliability, durability) using language that tends to be subjective and sentiment-laden (Liu, Lee, and Srinivasan 2019). Moreover, the number of product attributes that relate to fit can be large and vary across consumers. Individual consumers may care about different subsets of such attributes or may value the same attributes differently. For example, in the context of a smartphone, suppose that a consumer cares a lot about compatibility of the phone with an obscure hands-free protocol of an older vehicle. If no previous buyer of that product had been interested in that exact product attribute, it is unlikely that any related information would be present, either in the product description or in the available product reviews. Q&A technology would enable that consumer to proactively ask a question about that rather obscure product feature and thus resolve their idiosyncratic fit uncertainty prior to purchase. In such cases, Q&As act as a complement to reviews by allowing consumers to decrease their fit uncertainty prior to purchase. This in turn leads to consumers purchasing products better suited to their needs and, thus, to fewer postpurchase regrets among those who purchase.

The following hypotheses summarize the key predictions of our conceptual framework:

> **H_1:** Q&As can help resolve prepurchase fit uncertainty. Thus, consumers are better matched to products they purchase, resulting in higher subsequent product ratings.
>
> **H_2:** The effect of Q&As is expected to be stronger for products that have higher fit uncertainty.
>
> **H_3:** The addition of Q&As should reduce negative reviews that arise due to fit concerns.

In the Web Appendix, we present a model that captures how the presence of informative Q&As affects consumer decision making and product ratings in settings with consumer fit uncertainty. Our model supports the intuitions offered by this framework by showing that, if answers are reliably correct, Q&As increase ratings by eliminating fit mismatch. Further, the positive effect of Q&A on product ratings is proportional to the product-level probability that purchasers will experience fit mismatch without Q&A.

Data and Descriptives

We obtained data from Bazaarvoice via the Wharton Consumer Analytics Initiative (https://wcai.wharton.upenn.edu). Bazaarvoice provides software that enables businesses to collect and display reviews and Q&As on their websites. Our data come from a U.K.-based big-box retailer (similar to Amazon) that uses Bazaarvoice software. The data cover two product categories of consumer durables (Technology and Home & Garden) and include all reviews and Q&As posted between 2009 and 2015. The two product categories are further subdivided into 755 subcategories such as Bedroom Furniture and Video Games. It is worth noting that, unlike hotels/restaurants for durables it is less likely for firms to improve product quality over a short period of time, thus mitigating concerns of product quality changes in response to reviews.

Overall, our data contain 37,853 unique product identifiers.[4] Out of these, 19,961 products do not have any user-generated content (possibly because they were newly introduced products at the time of data collection) and thus cannot be considered for our analyses. Out of the remaining 17,892 products, 13,354 have at least one Q&A, 13,104 have at least one review, and 8,428 have both reviews and Q&As. Because we want to study the impact of Q&As on product ratings, our analyses focus on products that (1) have both reviews and Q&As and (2) have received at least one review before the first question was asked. This leaves us with 5,077 products, 345,168 reviews, and 48,687 Q&A pairs. Table 1 presents summary statistics for all products in our estimation.

In addition, we make use of click-stream data collected over a two-month period in 2015 (February and March) to supplement our main analyses. These data consist of customer browsing behavior (i.e., which specific product pages were clicked on). Within this data set, we look at products that received their first answer within the two-month observation period.

We also collected data from the Internet Archive to conduct a series of robustness checks. These supplementary data sets are described in detail in the "Robustness Checks" section.

[4] Some of these product identifiers refer to minor variations of the same underlying product, such as a white iPad and a black iPad, and share the same Q&As but have different reviews. We treat variations as independent products, because the ability of Q&A to alleviate fit uncertainty might differ across product variations and aggregating them would lead us to underestimate the treatment effect of interest. In addition, allowing for a product-level rather than a product-group-level fixed effect in our estimation lends more flexibility to the model.

Table 1. Summary Statistics.

	Products	Avg. Rating	Reviews	Questions
Technology				
Products with both Q&A and reviews	1175	4.37 (.55)	153.6 (138.5)	23.78 (34.84)
Home & Garden				
Products with both Q&A and reviews	3902	4.28 (.436)	264.5 (242.9)	14.2 (16.3)

Q&As, Reviews, and Fit Uncertainty

Our hypotheses relate to the ability of Q&As to resolve fit uncertainty. In this subsection, we show that the Q&A corpus has several characteristics that make it particularly well-suited to conveying information about fit to consumers. We also compare Q&As with consumer reviews and show that the two corpora differ in important ways that make Q&As better suited to resolving fit uncertainty.

We begin by examining the adoption of the Q&A technology, because the ability of Q&As to resolve fit uncertainty depends on the rate at which consumers use this feature. Figure 1 provides some descriptive patterns of Q&A dynamics. In Panels A and B, we find evidence of increased usage of Q&As over time, mirroring the increasing adoption of reviews. In Panel C, we show that, over time, questions have been getting answered faster: the average number of days it takes for a posted question to be answered has decreased from 17 days in 2011 to 4 days in 2015, suggesting increasing engagement with the feature. Finally, in Panel D, we plot the distribution of answers per question. In our data, all questions are answered, and approximately 70% of questions have a single answer. We also find that close to 80% of answers to questions come from the platform itself and not from other customers.[5]

In Table A1 in the Appendix, we display the top product categories in terms of questions per product. We find that categories related to electronics and their accessories receive the most questions per product. Because these products are complex and typically associated with customer concerns about usability and compatibility, we would indeed expect a larger number of questions related to them.

Next, we report some descriptive evidence consistent with our hypothesis that Q&As mostly contain prepurchase fit-specific information and do so in a less sentiment-laden fashion than reviews. First, to get at the prepurchase nature of Q&As, we randomly sampled 2,400 questions. A set of 240 coders were then asked to classify a randomly chosen set of ten questions each. We asked coders whether the question was most likely asked before or after purchase.[6] We then computed (at the coder level) the fraction of responses that were in favor of questions being before purchase. We find that 83% of questions are posted prepurchase. This is in stark contrast with reviews, which occur almost always (and for some platforms, exclusively) postpurchase.

To better understand content differences between reviews and Q&As, we perform a comparative sentiment analysis of the two corpora. We begin with a parts-of-speech classification of all reviews and Q&As. We find that reviews have a significantly higher proportion of adjectives and adverbs (20%) compared with Q&As (9%). Adjectives and adverbs are known to be important components of sentiment analysis (Benamara et al. 2007). We also perform a sentiment analysis task on Amazon Mechanical Turk by asking coders to rate the sentiment content of 2,000 Q&A pairs (each pair is rated by two independent coders, with a third coder being assigned to break any ties) and find that 90% of them are rated as "neutral." This leads us to believe that, although reviews are a more holistic expression of preferences, Q&As embody fit-specific information in a relatively sentiment-free way.

Finally, we examine the text of Q&As and reviews to gather additional evidence that Q&As are predominantly used for alleviating specific concerns related to product fit. To do this, we use a latent Dirichlet allocation (LDA) topic model (Blei, Ng, and Jordan 2003).[7] We train our LDA model on the entire body of reviews and Q&As and obtain 20 topics in each case. Consistent with our sentiment analysis, the topics obtained for Q&As contain more references to fit-related attributes (e.g., dimensions, compatibility) than the topics obtained for reviews, which mostly contain information about product quality or express sentiment in general. The top-three topics (and associated highest-probability words) obtained in both cases are provided in Table 2.[8]

Empirical Strategy

In this section, we begin by motivating our main estimating equation, and then discuss our identification strategy for measuring the causal impact of Q&As on consumer reviews.

Our main hypothesis is that Q&As can provide fit-specific information that allows for better-matched purchases, resulting in higher postpurchase utility and, thus, ratings. In our data, we observe individual reviews i left by consumers k of products j. In what follows, we are interested in modeling the utility U_{ij} derived by the i th purchase of product j regardless of the

[5] This may not be the norm across different e-commerce platforms. Anecdotally, Amazon seems to attract more customer answers. Future work could examine potential differences in the effects of Q&As in environments dominated by customer versus platform answers.

[6] For example, a prepurchase question would be "Is this keyboard compatible with MAC OS X Yosemite?," whereas a postpurchase question would be "My keyboard came with no instructions and the piece that raises the base already attached. How do I take it off?"

[7] We also build a naive Bayes classifier that discriminates between Q&A and review text and reach very similar qualitative conclusions: some of the top words that discriminate review content are "looks," "value," "money," and "great," whereas those for Q&As are "helps," "hope," "confirm," and "using." More details on this analysis are available upon request.

[8] For the complete set of topics, see the Web Appendix.

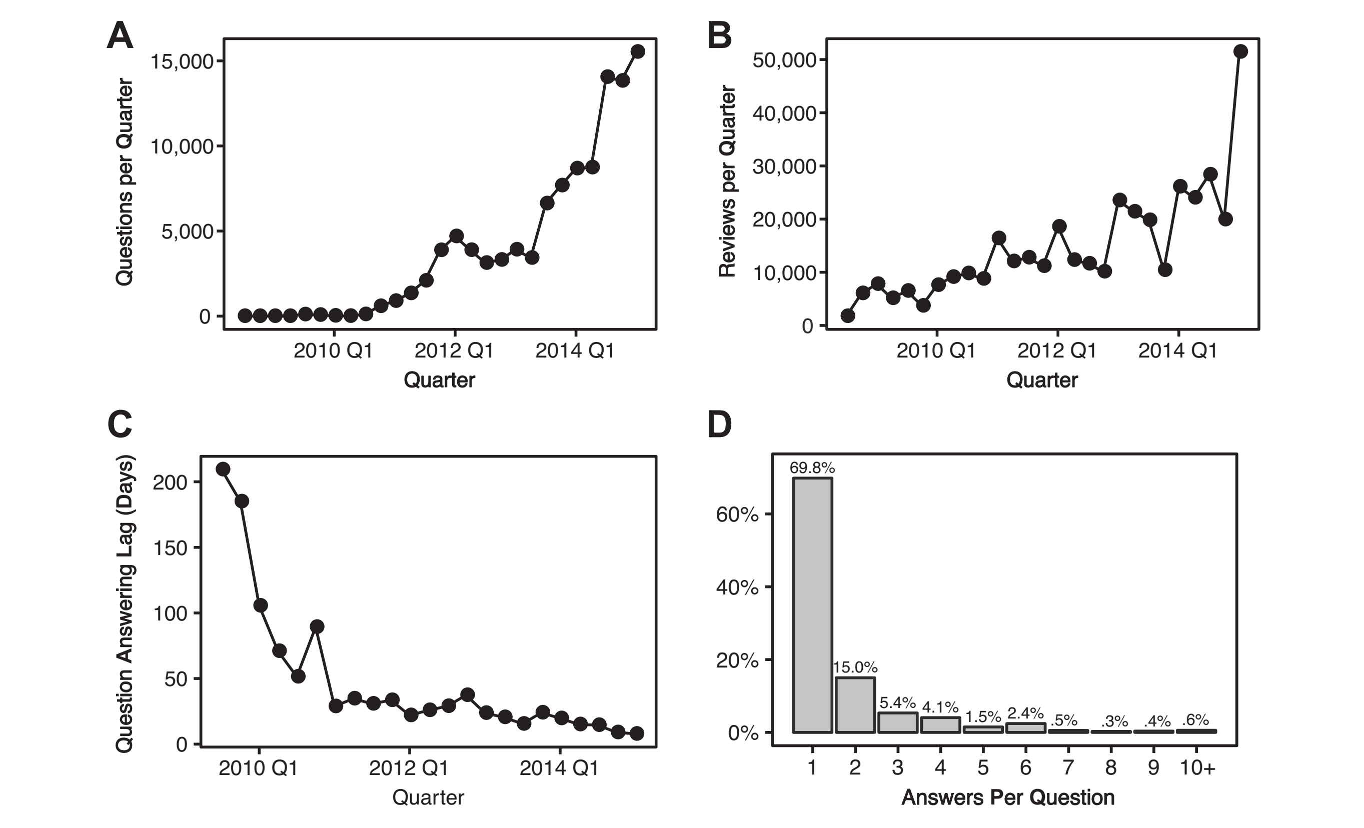

Figure 1. Descriptive features of user-generated content.

Notes: We find that the accumulation of questions has risen steadily over time (in parallel with reviews), questions have been attracting answers faster over time, and most questions have a single answer.

Table 2. Top-Three LDA Topics for Reviews and Q&As.

Reviews	Q&As
Quality (of Vacuum)	**Dimensions**
easy, great, good, cleaner, clean, product, vacuum	height, width, depth, dimensions, length, size, measurements
Quality (of Electronics)	**Guarantee/Warranty**
sound, good, great, clock, quality, set, radio	buy, bought, guarantee, product, warranty, year, item
Quality (of Phone/Camera)	**Compatibility (Computers)**
phone, easy, set, good, camera, features, box	ipod, compatible, laptop, windows, work, download, touch

identity of the consumer responsible for that purchase. This distinction between purchases i and consumers k is important because our theory predicts that the specific consumers who choose to buy a product under Q&A are different than the consumers who may buy the same product without Q&A.

We assume that products are both vertically and horizontally differentiated. Thus, each purchase i is associated with postpurchase utility from two separate components of product j: vertical quality (η_j), which is product specific and common to all purchasers, and horizontal fit (μ_{ij}), which captures the degree to which j is a good match for the consumer responsible for purchase i. Thus, ex post utility for a given purchase takes the following form:

$$U_{ij} = \eta_j + \mu_{ij}. \tag{1}$$

To capture the potential impact of Q&As, we model the horizontal component of postpurchase utility as

$$\mu_{ij} = \beta_j \times \mathrm{POST}_{ij} + \varepsilon_{ij}, \tag{2}$$

where POST_{ij} is an indicator set to one following each product's first answered question and zero otherwise, and ε_{ij} captures unobserved idiosyncratic factors that affect utility. To account for the fact that some products may suffer from fit mismatch more than others, we allow the effect of Q&As to be product-specific:

$$\beta_j = \beta_0 + \beta_1 \times m_j, \tag{3}$$

where m_j captures the degree of fit mismatch faced by purchasers of product j. Here, β_0 is an intercept term capturing the effect of Q&As on products facing no fit mismatch, and β_1 is a slope term capturing the effect of Q&As as the degree of mismatch m_j increases.

Substituting Equation 3 into the utility function in Equation 1, we derive the following expression for the average utility obtained by purchasers of product j:

$$U_{ij} = \eta_j + \beta_0 \times POST_{ij} + \beta_1 \times (m_j \times POST_{ij}) + \varepsilon_{ij}. \quad (4)$$

If β_1 is positive, we can infer that, on average, purchaser i is better matched with product j and therefore derives higher utility in the presence of Q&As. Further, the higher the degree of fit mismatch m_j, the greater the increase in utility. Under the standard assumption that reported ratings r_{ij} are an increasing function of experienced utility,[9] we modify Equation 4 to arrive at the following estimating equation:

$$r_{ij} = \eta_j + \delta_{t(ij)} + \beta_0 \times POST_{ij} + \beta_1 \times (m_j \times POST_{ij}) + X'_{ij} \times \beta_2 + \varepsilon_{ij}. \quad (5)$$

Compared with the utility function, Equation 5 introduces additional controls, some of which depend on calendar time. We use the subscript t(ij) to denote the year-month of review i for product j. Specifically, we model rating i left for product j as a function of product fixed effects η_j, time fixed effects $\delta_{t(ij)}$, time-varying observables X_{ij}, the $POST_{ij}$ indicator, the fit mismatch term m_j, and unobservables ε_{ij}. In all specifications we estimate, we cluster standard errors at the product level (Donald and Lang 2007).

The coefficient of interest, β_1, has a causal interpretation under the assumption that the timing of each product's first answer is as good as random. This assumption will be violated if increases in ratings and the propensity to answer questions are jointly driven by an unobserved process. In the "Robustness Checks" section, we discuss and mitigate specific threats to this assumption.

Measuring Fit Mismatch

Our theoretical framework predicts that the effect of Q&A on product ratings is proportional to the degree of fit mismatch (m_j) that purchasers will experience without Q&A. However, the degree of fit mismatch inherent in any product's past purchases is not directly observable, because we do not know which negative reviews arise due to poor quality, fit mismatch, or other reasons. We approach this problem by constructing three proxies for the presence of fit mismatch prior to Q&A, which we describe next.

Low average ratings. Products with a high pre-Q&A probability of fit mismatch will be purchased by many consumers for whom the product is a poor fit. In turn, these consumers will leave negative reviews for these products, leading to low average ratings. Thus, a simple heuristic for identifying products that suffer from bad fit is to focus on products with low average ratings prior to treatment (i.e., before the arrival of the first answer). We use an average star rating of four out of five as the threshold that separates these products that may suffer from fit mismatch from those that do not. Subsequently, we also show that our results are robust to different thresholds. The main concern with this measure is that fit mismatch is not the only source of negative reviews. Thus, by selecting products that have low average ratings prior to receiving Q&As, we may also incorrectly include products that do not suffer from fit mismatch (e.g., low-quality products). These false positives may attenuate the average treatment effect on the treated we estimate.

High rating variance. Our second measure examines products that have a high rating variance (Sun 2012). We can think of such products as suffering from fit mismatch because they have attributes that are asymmetrically preferred by consumers (i.e., some like them and find them to be a good fit, whereas others do not). We label products whose rating variance is greater than 1 (the median) as suffering from mismatch. This measure also runs the risk of attenuation bias, albeit in a different sense: for products that have both high variance and high average ratings, bad fit might not be a dominant concern, and therefore, Q&As might have less of an impact.

Thus, both the low-ratings and high-variance measures can misclassify products as suffering from poor fit when they do not. For example, when both the mean and variance of ratings are low, the most likely cause is poor quality rather than poor fit. Drawing on this observation, we also estimate specifications that combine these two proxies to identify products suffering from poor fit. Our expectation is that Q&As will be particularly helpful for low-rated, high-variance products.

Review text. Our final measure looks at review text to identify products suffering from fit mismatch. To construct this measure, we build a text classifier that can distinguish negative reviews that arise due to poor fit. Using the classifier, we label each negative review as fit-related or not. Finally, we construct a continuous variable for fit mismatch for each product as the fraction of negative fit related reviews prior to each product's first answer.

To build the classifier, we first construct a training set by asking two coders (on Amazon Mechanical Turk) to select the most likely cause of 3,300 randomly chosen negative (one-, two-, and three-star) reviews. We indicate three categories into which reviews are to be classified: poor fit, poor quality, or other miscellaneous reasons (e.g., issues with the store or shipping). The Web Appendix shows the survey seen by the coders. Any disagreements were resolved by a third coder. The coders classified 28% of negative reviews as having resulted primarily from poor fit and 67% primarily from poor quality. Because the third category accounted for a small fraction of the reviews (<5%), relating mostly to in-store experiences and returns, we ignore it in our subsequent analysis.

We use these manually labeled reviews to train a C-support vector machine (C-SVM) classifier (Cortes and Vapnik 1995). To perform the classification task, we remove common stop words and then tokenize and stem the text of each negative review into a bag-of-words representation, thus

[9] This includes, for example, situations in which consumers report their utility relative to their expectations (Hu, Pavlou, and Zhang 2017) in their ratings and allows for the fact that only a subset of purchasers leave a rating.

Table 3. Top-Five Reviews with the Highest Predicted Probabilities of Quality and Fit Issues.

Quality	Fit
"This fan was not worth the money. It is poor quality for value. wasnt very efficient and broke after a couple of weeks."	"The boxes are small but quite strong. Quite small for the price"
"poor quality, some screws missing."	"Good quality—but too small for my requirement would have prefered it bigger"
"Poor quality, flimsy, and parts missing. Returned!"	"I use as back up to ecomy 7 it is a bit small though should have got two or a bigger one"
"This clothes dryer fell apart before I had even erected it. Very flimsy and poor quality. Took back next day!"	"Bit small for what I needed if for"
"Poor quality, ended up in bin."	"Good Figures too small"

obtaining word frequencies for each negative review. We use these word frequencies as predictors to train a classifier that predicts whether a negative review arises primarily from poor fit or poor quality.[10]

We train our classifier on an 80% random sample of our labeled data, holding out the remaining 20% to evaluate the classifier's performance. The C-SVM classifier has one tunable parameter, C, which intuitively calibrates the trade-off between classification accuracy and a larger-margin separating hyperplane. We select a value for C using fivefold grid search cross-validation. We evaluate the out-of-sample performance of our classifier using the commonly employed receiver operating characteristic area under the curve (ROC-AUC) metric. ROC-AUC ranges from 0 to 1 and is a ranking metric. Intuitively, an ROC-AUC value of p implies a p probability of correctly predicting which of two reviews belonging to different classes (poor fit and poor quality) belongs to the poor-fit class. Our classifier achieves an ROC-AUC of .82.[11]

In addition to assessing the out-of-sample predictive power of our classifier, we also check whether our classifier makes qualitatively meaningful predictions about fit mismatch. We do so in three ways. First, in Table 3 we present the top five reviews with the highest predicted probability of belonging to each of the two classes (fit vs. quality issues). While reviews with a high predictive probability of being about quality issues are explicit in mentioning poor product performance, reviews identified as having fit issues reflect customer-specific requirements that the product failed to fulfill, despite not inherently being of an inferior quality.

Second, we order all product categories in our data set by the fraction of negative reviews that arise due to fit issues. We present these results in Figure 2. Intuitively, we would expect that categories for which a higher fraction of negative reviews are about fit would tend to be those for which it is harder for consumers to gauge whether the product is right for them. Indeed, we find that sofas (for which look and feel might be difficult to gauge), electronic devices (which may involve compatibility issues) or accessories of some kind (which are meant to supplement a diverse set of existing items) tend to have more negative fit reviews. In contrast, products for which the customer's domain knowledge dictates their purchase (e.g., do-it-yourself purchases, power tools) have fewer fit concerns according to our classifier.

Third, we examine the top 20 words that are most predictive of fit versus quality issues. To do so, we use layer-wise relevance propagation (LRP), a method originally developed to interpret the results of deep neural nets (Bach et al. 2015). LRP produces a score for each word and class (poor fit and poor quality). High scores are assigned to words that are good at discriminating reviews belonging to each class. For linear SVMs, the LRP score of each word–class combination is computed as the sum of the products of the word loading and the term frequency–inverse document frequency score of the word in each of the reviews belonging to that class. We present the top-ranking words by LRP score for each of the classes in Table 4. We find, as expected, that words that are a measure of objective quality (work, poor, cheap, broke) tend to be more predictive of negative quality reviews, whereas words that indicate more person-specific, idiosyncratic attributes (look, need, design, however) are predictive of negative fit reviews.

Overall, these results suggest that our text classifier can discriminate between reviews that bring up fit-related concerns and those that do not.

Mean Reversion and Measurement Error

A final empirical challenge arises due to the fact that we construct proxies for fit mismatch as a function of past ratings, or quantities correlated with past ratings (e.g., review text). This leads to two problems, which arise even if we assume treatment is strictly exogenous (i.e., $\mathbb{E}[POST_{ij} \times \varepsilon_{ij}] = 0$). Here, we discuss these two problems under the assumption of treatment exogeneity; we discuss violations to treatment exogeneity separately in the "Robustness Checks" section.

The first problem arises from applying a within transformation to Equation 5 to eliminate product fixed effects η_j. The transformation mechanically introduces correlation between the demeaned residual and the demeaned version of $m_j \times POST_{ij}$, violating strict exogeneity and biasing ordinary least squares (OLS) estimates of β_1. (To see this, note that demeaning $m_j \times POST_{ij}$ and ε_{ij} introduces the mean error term in both quantities.) Although this type of bias is more commonly seen in models that explicitly incorporate a lagged outcome as a control (Nickell 1981), it also arises in our setting because m_j is a function of lagged outcomes.

The second problem arises due to measurement error in the fit mismatch measure m_j. Recall that we do not observe m_j

[10] We also considered using bigrams as predictors, but we did not see significant improvement in out-of-sample predictive power.

[11] We also replicate our main analyses with a naive Bayes classifier, which achieves an ROC-AUC of .79. These results are available upon request.

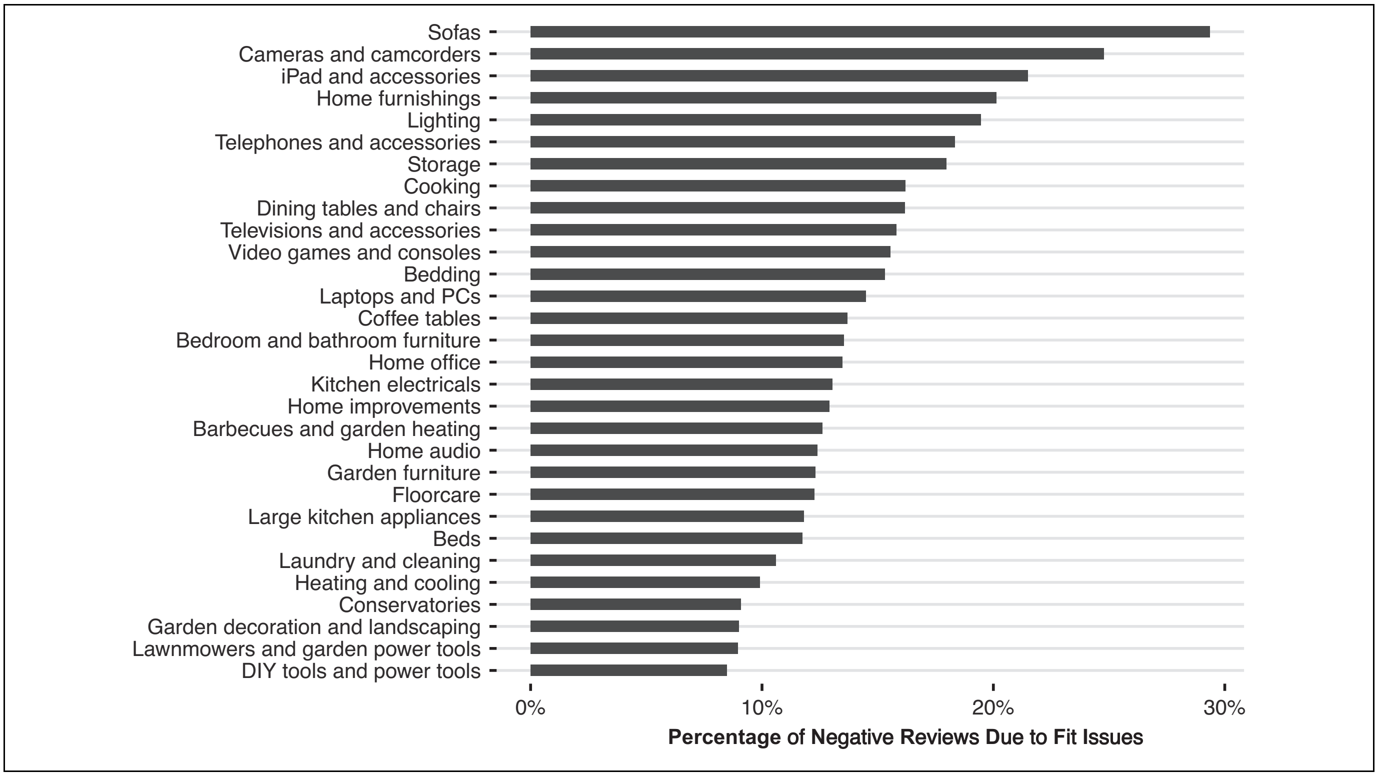

Figure 2. Percentage of negative reviews (less than or equal to three stars) that are due to fit-related issues by product category
Notes: We limit our analysis to categories with at least 100 products and at least 100 reviews.

Table 4. Top-20 Words Most Predictive of Quality/Fit Issues, Estimated Using Layer-Wise Relevance Propagation on the Output of an SVM Classifier.

Quality	Fit
work	small
quality	look
poor	need
very	was
cheap	size
flimsy	does
return	fit
buy	just
good	design
money	colour
broke	shelf
make	use
any	however
week	bigger
day	did
screw	suitable
miss	picture
try	difficult
got	comfort
open	differ

directly, instead relying on noisy measures $\widetilde{m}_j = m_j + v_j$ (where $\widetilde{m}_j$ in our case could be mean ratings or rating variance), and v_j reflects unobserved factors uncorrelated with m_j that enter these proxies. For instance, some products may randomly experience transient shipping delays—a random shock to v_j—prior to their first Q&A leading to excess negative ratings. This could decrease the products' mean ratings and increase rating variance (which we use as proxies for fit mismatch) for reasons unrelated to fit mismatch. Subsequent ratings for these products will likely revert back to their mean levels (e.g., once shipping delays are resolved) regardless of any direct Q&A effect.

Rewriting our main estimating equation to reflect the use of proxies for fit uncertainty $\widetilde{m}_j$, we have the following:

$$\begin{aligned} r_{ij} &= \eta_j + \delta_{t(ij)} + \beta_0 \times POST_{ij} + \beta_1 \times (m_j \times POST_{ij}) + X'_{ij} \\ &\quad \times \beta_2 + \varepsilon_{ij} = \eta_j + \delta_{t(ij)} + \beta_0 \times POST_{ij} + \beta_1 \\ &\quad \times \left[(\widetilde{m}_j - v_j) \times POST_{ij}\right] + X'_{ij} \times \beta_2 + \varepsilon_{ij} \\ &= \eta_j + \delta_{t(ij)} + \beta_0 \times POST_{ij} + \beta_1 \times (\widetilde{m}_j \times POST_{ij}) \\ &\quad + X'_{ij} \times \beta_2 + \widetilde{\varepsilon}_{ij}, \end{aligned} \tag{6}$$

where

$$\widetilde{\varepsilon}_{ij} = -\beta_1 \times (v_j \times POST_{ij}) + \varepsilon_{ij}. \tag{7}$$

Note that $Cov(\widetilde{m}_j \cdot POST_{ij}, \widetilde{\varepsilon}_{ij}) \neq 0$, because both quantities depend on the unobservable v_j. This results in bias when Equation 6 is estimated by OLS.

We adopt a standard solution (Griliches and Hausman 1986) to this classical measurement error problem, relying on a second noisy measure of our proxy, which we use as an instrument. Other articles that have used a similar empirical strategy in marketing include Narayanan and Nair (2013); it has also seen widespread use in economics (e.g., Acemoglu and Finkelstein 2008; Gupta 2017). Specifically, we divide the pre-Q&A period for each product into two smaller samples: a holdout period, which includes all reviews up to 200 days before the first answer, and a shorter pretreatment period that includes all reviews starting at 200 days before the first answer and ending at the time of the first answer.[12] We then use these two samples to construct the two proxies $\widetilde{m}_j^{holdout}$ and $\widetilde{m}_j^{pre}$, where the former quantity can be considered a lag of the latter. Finally, we use $\widetilde{m}_j^{holdout}$ to construct instruments for the endogenous variable $\widetilde{m}_j^{pre} \times POST_{ij}$. The holdout sample is subsequently excluded from estimation. Our main estimating equation and the corresponding first stage are given by

$$r_{ij} = \eta_j + \delta_{t(ij)} + \beta_0 \times POST_{ij} + \beta_1 \times (\widehat{\widetilde{m}_j^{pre} \times POST_{ij}}) + X'_{ij} \times \beta_2 + \widetilde{\varepsilon}_{ij}, \tag{8}$$

$$(\widetilde{m}_j^{pre} \times POST_{ij}) = \widetilde{\eta}_j + \widetilde{\delta}_{t(ij)} + \gamma_0 \times POST_{ij} + \gamma_1 \times (\widetilde{m}_j^{holdout} \times POST_{ij}) + X'_{ij} \times \gamma_2 + \widetilde{u}_{ij}, \tag{9}$$

where $\widetilde{\varepsilon}_{ij} = -\beta_1 \times (v_j^{pre} \times POST_{ij}) + \varepsilon_{ij}$ and $\widetilde{u}_{ij} = -\gamma_1 \times (v_j^{holdout} \times POST_{ij}) + u_{ij}$. To see why this strategy works, note that

$$\widetilde{m}_j^{holdout} = m_j + v_j^{holdout}, \tag{10}$$

$$\widetilde{m}_j^{pre} = m_j + v_j^{pre}. \tag{11}$$

The instrument $\widetilde{m}_j^{holdout} \times POST_{ij}$ is valid under two conditions. First, it has to be relevant and have a strong first stage, which we can verify. Second, it has to satisfy the exclusion restriction $\mathbb{E}\left[(\widetilde{m}_j^{holdout} \times POST_{ij})\widetilde{\varepsilon}_{ij}\right] = 0$. This condition will be met as long as the two measurement errors are not correlated, that is, $\mathbb{E}\left[v_j^{holdout} \times v_j^{pre}\right] = 0$ (recall that we are assuming ε_{ij} is otherwise exogenous).

Intuitively, and continuing our prior example, we are assuming that products that experienced random shipping delays (and consequently excess negative ratings) in the pretreatment period were not more likely to also experience such shocks in the holdout period. Thus, by instrumenting with holdout ratings we use the signal embedded in holdout measures of fit mismatch (m_j) to get rid of the noise in pretreatment measures of fit mismatch (v_j^{pre}), the latter being the source of bias when we estimate Equation 6 by OLS.

Although we cannot directly test the exclusion restriction, we check whether ratings, which we use to construct proxies for fit mismatch, are serially correlated conditional on observables. To do so, we conduct an autocorrelation test proposed by Arellano and Bond (1991; Roodman 2009), and find that autocorrelation in levels vanishes beyond the first lag. Specifically, a rating may be serially correlated with the rating directly preceding it, but this serial correlation decays fast and is not statistically significant for the second lag and beyond. This provides us with some confidence that functions of ratings that are far apart in time (such as $v_j^{holdout}$ and v_j^{pre}) are not correlated.

Using the best linear unbiased predictor to construct instruments. As we have discussed, the proxies we use for fit mismatch are measured with error. Beyond causing issues with identification, measurement error means that the instrument $\widetilde{m}_j^{holdout} \times POST_{ij}$ may be a poor predictor of $\widetilde{m}_j^{pre} \times POST_{ij}$ for products with few reviews in the holdout period. Here, we explain how we obtain more precise measurements of our fit mismatch measures.

To obtain a stronger instrument, we use a shrinkage estimator for $\widetilde{m}_j^{holdout}$, where we shrink the estimates of $\widetilde{m}_j^{holdout}$ for products with few reviews toward the population mean (Robinson 1991). Specifically, for each product we estimate the best linear unbiased predictor (BLUP) of its mean rating in the holdout sample using a mixed-effects model with a random intercept m_j for each product j, and a fixed intercept μ:

$$\widetilde{m}_j^{holdout} = \mu + m_j + e_j.$$

We apply this shrinkage estimator to the average rating and fraction of fit-related negative reviews instruments. Estimating the above equation, we obtain a BLUP of the mean rating and the mean fraction of fit related negative reviews in the holdout sample for each product, which we use to construct our final instruments.[13]

Results

Next, we estimate the effect of Q&A arrival on ratings using each of the three fit mismatch measures described previously.

Low Pre-Q&A Ratings

Our first measure is low pretreatment average ratings. Thus, we estimate Equation 8 with $\widetilde{m}_j^{pre}$ being an indicator for products with low average pretreatment ratings (less than or equal to four stars), and $\widetilde{m}_j^{holdout}$ being the average rating of the product in the holdout period. In addition to product and time fixed effects, we also control for the rank of each review (as recommended by, e.g., Godes and Silva [2012]).

We present our results in Table 5. Column 1, which shows our OLS estimates, serves as the baseline and includes all products and all reviews in the estimation sample. We find a

[12] In a robustness check, we change the definition of our holdout sample to make it more flexible: out of all pre-Q&A observations, we select the most recent 50% to form the pretreatment sample, and the rest to form the holdout sample. We report these estimates, which are similar to our main results, in the Web Appendix.

[13] Results remain qualitatively unchanged even without employing the shrinkage estimator. However, they are slightly attenuated due to higher measurement error in m_j for products with fewer reviews. These results are available upon request.

Table 5. The Impact of Q&A on Ratings Using Low Pretreatment Ratings (Less Than or Equal to Four Stars) as a Measure for Fit Mismatch.

	OLS (Full Sample)	OLS (Excl. Holdout)	IV (First Stage)	IV (Second Stage)	IV (Second Stage, Bins)
POST × Low rating	.243***	.206***		.122***	
	(.022)	(.031)		(.041)	
POST × Holdout rating			−.901***		
			(.029)		
POST	−.045***	−.026**	4.049***	−.011	
	(.009)	(.012)	(.130)	(.012)	
POST × Rating ∈ [2, 3]					.502***
					(.101)
POST × Rating ∈ (3, 4]					.100***
					(.037)
POST × Rating ∈ (4, 5]					−.012
					(.012)
Review rank	.0002***	.0001*	−.0001*	.0001*	.0001*
	(.0001)	(.0001)	(.00004)	(.0001)	(.0001)
Product FEs	Yes	Yes	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes	Yes	Yes
F-statistic			312.91		
Observations	345,168	184,811	184,811	184,811	184,811

*$p < .1$.
**$p < .05$.
***$p < .01$.
Notes: FEs = fixed effects. Standard errors are clustered at the product level.

positive and significant increase in ratings of .24 for products with a low pretreatment mean. We also find a statistically significant decrease of .045 stars for products that have a high pretreatment mean. As discussed in the "Mean Reversion and Measurement Error" subsection, some products may have high or low pretreatment ratings by pure chance rather than as a consequence of poor fit. We would expect the ratings of these products to mean-revert regardless of Q&As, which would inflate our estimates. In column 2, we reestimate the OLS model by excluding the holdout sample. Now, we see that both effects decrease in magnitude, but we still observe a small dip for products without fit uncertainty. Next, we move on to the instrumental variables (IV) specification. Column 3 reports the first stage of Equation 8. We see that average ratings computed on the basis of the holdout sample using BLUP are strong predictors of the pretreatment average rating. Column 4 reports our preferred IV estimate: we find a positive and significant increase of .12 for low-rated products and see no corresponding decrease for high-rated products. Finally, in column 5, to capture treatment heterogeneity, we estimate a less parsimonious but more flexible specification where we interact the $POST_{ij}$ variable with a full set of dummies for m_j being in different unit intervals (thus, we do not include the main effect for $POST_{ij}$). We instrument each of these variables with the corresponding lagged version from the holdout sample. We find substantial heterogeneity: the higher the fit mismatch, the larger a product's posttreatment increase in ratings.

To put the magnitude of our effect—approximately .1 stars on average—in context, we compare it with the standard deviation of average ratings, which, for products with at least five reviews, is also roughly equal to .1 stars. The size of the effect we estimate is comparable to that of similar interventions such as the adoption of management responses (Proserpio and Zervas 2017).

High Pre-Q&A Variance

Our second measure of fit mismatch is high rating variance prior to treatment. We estimate Equation 8 with $\widetilde{m}_j^{pre}$ being an indicator for products with pretreatment variance ≥ 1 (the median value) and $\widetilde{m}_j^{holdout}$ being the rating variance in the holdout period. We report our results in Table 6 for the full sample OLS (column 1), OLS excluding holdout data (column 2), and IV (columns 3 and 4). In all cases, we find a positive and significant increase in ratings for high-variance products following treatment. Similar to the previous case, we observe a small decrease for products with low variance, but this effect is very close to zero once we instrument for mean reversion.

As described in the "Measuring Fit Mismatch" subsection, we also estimate specifications that interact the low-ratings and high-variance proxies to better identify the set of products whose ratings have suffered due to fit-related concerns. This specification contains the full interaction between the rating and variance proxies resulting in four groups of products based on their pretreatment ratings: low rating/low variance, high rating/low variance, low rating/high variance, and high rating/high variance. We instrument each of these dummies with its holdout equivalent. Given our theory, we expect low-rating/high-variance products to be primarily impacted by Q&As. We present our results in Table 7. Among the four groups of products, we see a

Table 6. The Impact of Q&A on Ratings Using High Pretreatment Rating Variance (Greater Than or Equal to One Star) as a Measure for Fit Mismatch.

	OLS (Full Sample)	OLS (Excl. Holdout)	IV (First Stage)	IV (Second Stage)
POST × High variance	.159***	.150***		.085***
	(.015)	(.019)		(.030)
POST × Holdout variance			.533***	
			(.016)	
POST	−.061***	−.053***	−.064***	−.026*
	(.010)	(.012)	(.019)	(.015)
Review rank	.0001***	.0001	.00002	.0001*
	(.00004)	(.0001)	(.00004)	(.0001)
Product FEs	Yes	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes	Yes
F-statistic			380.08	
Observations	345,168	184,811	184,811	184,811

*$p < .1$.
**$p < .05$.
***$p < .01$.
Notes: FEs = fixed effects. Standard errors are clustered at the product level.

Table 7. The Impact of Q&A on Ratings Using Both Low Pretreatment Ratings (Less Than or Equal to Four Stars) and High Pretreatment Rating Variance (Greater Than or Equal to One Star) as Measures for Fit Mismatch.

	IV (Second Stage)
POST × Low rating × Low variance	−.594
	(.429)
POST × Low rating × High variance	.125***
	(.038)
POST × High rating × Low variance	−.010
	(.014)
POST × High rating × High variance	−.013
	(.028)
Review rank	.0001*
	(.0001)
Product FEs	Yes
Year-month FEs	Yes
Observations	184,811

*$p < .1$.
**$p < .05$.
***$p < .01$.
Notes: FEs = fixed effects. Standard errors are clustered at the product level.

statistically significant increase in ratings following Q&A arrival only for the low-rating/high-variance group, as expected. It is worth noting that we do not see a significant effect for high-rating/high-variance products. As previously mentioned, fit mismatch might not be the dominant concern for this group of products, because they are already high-rated.

High Proportion of Pre-Q&A Fit-Related Negative Reviews

Our final measure uses review text to detect pre-Q&A concerns about fit that might exist for a product. As a measure of the probability of fit mismatch, we compute the fraction of all reviews that are negative and fit-related prior to the arrival of the first question, based on the classifier described in the "Measuring Fit Mismatch" subsection. We estimate Equation 8 with $\widetilde{\mathrm{m}}_{\mathrm{j}}^{\mathrm{pre}}$ being the fraction of pretreatment fit-related negative reviews and $\widetilde{\mathrm{m}}_{\mathrm{j}}^{\mathrm{holdout}}$ being the fraction of fit-related negative reviews constructed in the holdout sample estimated using a logistic BLUP. We present our results in Table 8. The effect sizes for each specification mirror those found previously, thus indicating that the extent of rating increase is proportional to the fraction of fit related negative reviews. To illustrate how the effect sizes can be interpreted, consider the estimate in column 4: for a product with 10% of pretreatment reviews expressing fit concerns, we estimate a subsequent increase in ratings of $1.135 \times .1 = .11$ stars.

Mechanism: Fit Mismatch Reduction

Next, we examine a hypothesized mechanism for our effect—namely, that Q&As lead to higher ratings by promoting better matches between consumers and products. To do so, we estimate the same specification as Equation 5, but with the dependent variable being an indicator for fit-related negative reviews (based on the classifier described in the "Measuring Fit Mismatch" subsection). We code all non-fit-related negative reviews and all positive reviews (four and five stars) as zero. To match the definition of our dependent variable, we use the text-based measure for fit mismatch. (We obtain similar results for our two other measures, low ratings and high variance.)

We present results for both OLS and IV specifications in Table 9 and include a control for review rank. We find a negative and significant effect for the impact of Q&As on the probability of receiving a negative fit-related review for each of our specifications, indicating that the fraction of fit-related reviews declines following the arrival of the first answer, in proportion to the degree of fit mismatch prior to Q&A. Focusing on our preferred IV specification in column 4, we can interpret our

Table 8. The Impact of Q&A on Ratings Using the Pretreatment Fraction of Reviews Mentioning Fit Issues as a Measure for Fit Mismatch.

	OLS (Full Sample)	OLS (Excl. Holdout)	IV (First Stage)	IV (Second Stage)
POST × Fit	1.932***	1.948***		1.135***
	(.150)	(.288)		(.356)
POST × Holdout fit			1.452***	
			(.054)	
POST	−.039***	−.030**	−.007***	−.013
	(.010)	(.012)	(.001)	(.013)
Review rank	.0002***	.0001**	−.00000**	.0001**
	(.00005)	(.0001)	(.00000)	(.0001)
Product FEs	Yes	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes	Yes
F-statistic			239.46	
Observations	345,168	184,811	184,811	184,811

*$p < .1$.
**$p < .05$.
***$p < .01$.
Notes: FEs = fixed effects. Standard errors are clustered at the product level.

Table 9. Mechanism: Products with a High Fraction of Pretreatment Fit-Related Negative Reviews Experience a Decline in Such Reviews Following Q&A.

	OLS (Full Sample)	OLS (Excl. Holdout)	IV (First Stage)	IV (Second Stage)
POST × Fit	−.726***	−.590***		−.190***
	(.024)	(.065)		(.072)
POST × Holdout fit			1.452***	
			(.054)	
POST	.016***	.013***	−.007***	.005***
	(.001)	(.001)	(.001)	(.002)
Review rank	−.00002***	−.00001	−.00000**	−.00000
	(.00000)	(.00000)	(.00000)	(.00000)
Product FEs	Yes	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes	Yes
F-statistic			239.46	
Observations	345,168	184,811	184,811	184,811

*$p < .1$.
**$p < .05$.
***$p < .01$.
Notes: FEs = fixed effects. Standard errors are clustered at the product level.

estimates as follows: if 10% of all pretreatment reviews are due to fit uncertainty, the product would experience a subsequent decline in the probability of receiving a negative fit-related review of $-.19 \times .1$ (i.e., 1.9%). This effect is relatively small due to the fact that the probability of receiving a negative review is low to begin with: in our data, only 15% of all reviews are negative (one, two, and three stars.) However, conditional on receiving a negative review post-Q&A, the probability that this review is fit-related decreases by $100/15 \times 1.9 = 12.6\%$.

Here, we have shown that Q&As lead to fewer negative reviews that contain fit-related concerns. Our hypothesized theory for this reduction is that Q&As change the mix of consumers who purchase a product (i.e., Q&As affect selection into purchasing a product by helping consumers discover whether a product is a good match for them). However, Q&As might also affect who decides to leave a review. For instance, some consumers may make mismatched purchases because they neglected to read Q&As addressing their fit concerns. These consumers may later avoid leaving negative reviews if they realize that the poor purchase was their own mistake. However, we believe that this is unlikely to be the prevalent mechanism for two reasons. First, it assumes that consumers who did not read Q&As when they were researching a product decided to read them before leaving a review. While this is possible, we think it is unlikely. Second, if Q&As deter people from leaving a review, we might expect to see a reduction in the volume of reviews post Q&A arrival. However, this is not what we find (see Table 10).

Robustness Checks

Our results indicate that answering a question leads to an increase in subsequent ratings for products that have suffered the consequences of fit mismatch. Moreover, we show that this

Table 10. Review Volume Around the First Answer.

	IV (Second Stage)	IV (Second Stage)	IV (Second Stage)
POST × Low rating	−.033		
	(.071)		
POST × High variance		.020	
		(.052)	
POST × Fit			−.543
			(.732)
POST	.094	.080	.100
	(.075)	(.078)	(.076)
Product FEs	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes
Weekday FEs	Yes	Yes	Yes
Observations	25,257	25,257	25,257

Notes: FEs = fixed effects. Standard errors are clustered at the product level.

Table 11. Page Views Around the First Answer.

	IV (Second Stage)	IV (Second Stage)	IV (Second Stage)
POST × Low rating	3.708		
	(4.191)		
POST × High variance		1.413	
		(1.133)	
POST × Fit			7.216
			(9.394)
POST	−.987	−.632	−.186
	(1.630)	(1.040)	(.819)
Product FEs	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes
Weekday FEs	Yes	Yes	Yes
Observations	8,614	8,614	8,614

Notes: FEs = fixed effects. Standard errors are clustered at the product level.

increase is driven by fewer fit-related negative reviews post-Q&A. The primary threat to these findings is an unobserved time-varying confounder that drives both the arrival of Q&As and a subsequent increase in ratings, at any time in the post-period, for products that suffer from mismatch. In this section, we investigate three such plausible confounders.

The first confounder we consider is promotions and discounts. Both increased advertising and reduced prices can increase demand for a product, resulting in more questions being asked and more reviews being submitted. The ratings associated with these new reviews may be higher than the product's current average rating due to lower prices or may be due to a well-targeted advertising campaign that drives purchases from consumers who are likely to enjoy the product.

Next, we consider improvements to the product page. In response to a question being asked, the platform may update a product's description, which could alleviate fit uncertainty and thus increase ratings. In this scenario, while Q&A spurs the improvement of the product page, it is not the direct cause of increased ratings.

Finally, we consider an influx of fit-related reviews just prior to treatment. Here, it would be these new reviews that help consumers discover products that are better suited to their needs rather the Q&A.

We address these concerns through a series of robustness checks. First, we show that there are no changes in review volume or product page views around each product's first answer, which we would expect in the presence of increased advertising. Next, we collect additional data that enable us to construct a panel of product descriptions, prices, and whether a product was being discounted. We find that our results are robust to controlling for price, discounts, and description lengths. Using the same data set, we also show that the content of product descriptions does not change significantly around the time the first Q&A arrives. Finally, we check whether the content of reviews changes around the first Q&A and find this not to be the case. We describe these robustness checks in detail in the following subsections.

Changes in Review Volume or Page Views

A product-specific marketing campaign could raise demand for the product, leading to more Q&As and, if the marketing campaign is well-targeted, higher ratings. To guard against this concern, we look for direct evidence that a marketing campaign may have taken place around the time of each products' first answer.[14] We focus on two outcomes suggestive of increased marketing activity: the daily number of reviews and the daily number of page views each product receives.

First, we examine whether review volume increases significantly following each product's first answer. To do so, we estimate Equation 8 using the daily count of reviews each product receives as the dependent variable. As before, we instrument for fit mismatch using holdout measures and include weekday fixed effects as additional controls. We present our results using each of the three fit mismatch measures in Table 10. We find no significant change in review volume around the first answer. This also hints at the fact that while Q&As improve match, they do not have a direct effect on demand (proxied by the number of reviews) in the short term.

One concern with this analysis is low power. Because reviews are relatively infrequent, a change in review volume can be difficult to detect. To increase power, we use clickstream data made available to us for a two-month period (February and March 2015) and repeat our analysis using daily page views—a more frequent event—as our outcome. We estimate our regression using 1,091 products that receive their first answers during that period. We present our results in Table 11. We see no significant change in page views post treatment.[15]

Finally, we graphically examine any changes in review volume or page views in the immediate neighborhood of the first

[14] We focus on a 180-day period centered on the first answer, but our results are robust to other windows of time.

[15] We further address the issue of low power by using OLS on the full sample (results reported in Tables A2 and A3 of the Appendix) and still find null effects. Arguably, estimating a precise zero effect even when mean reversion is present is a more stringent test of our hypothesis.

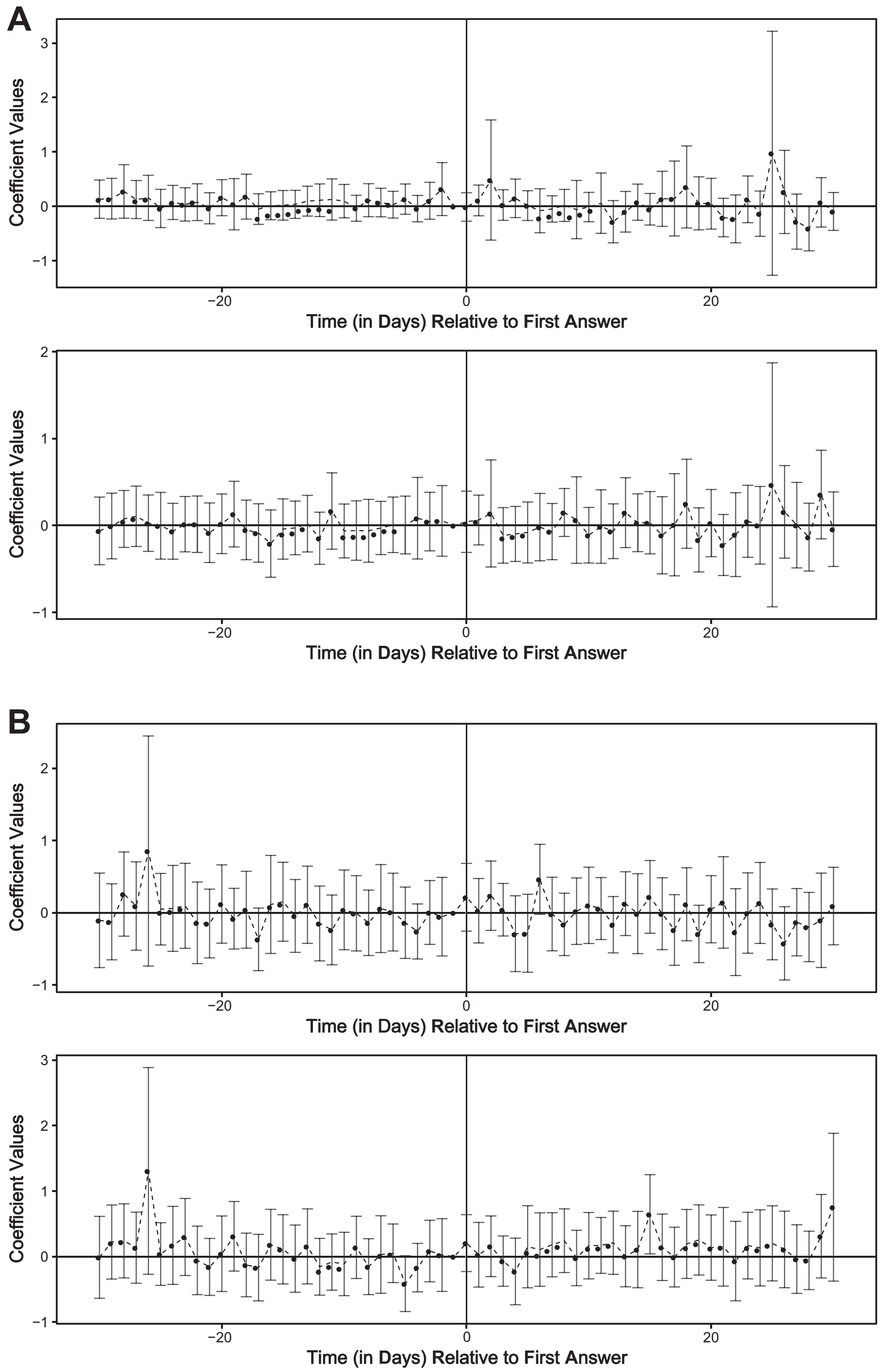

Figure 3. The evolution of review volume and page views for products with low ratings (top) and high variance (bottom), pre- and posttreatment (indicated by 0), measured 30 days around the first answer.

Notes: The points plot the β_k coefficient estimates from Equation 12, and the bars indicate the 95% confidence interval. We see that there are no significant irregularities in review volume or page views around the first answer time, thus mitigating the threat of omitted variables such as promotions.

answer. To do so, we estimate the following model for our two discrete measures of fit mismatch, low rating and high variance:

$$y_{jt} = \eta_j + \delta_t + \sum_{k=-30}^{30} \beta_k \times I\{D_{jt} = k\}_j + \varepsilon_{jt}, \qquad (12)$$

where y_{jt} is respectively the number of reviews or page views and $I\{D_{jt} = k\}$ is an indicator for day $k \in -30, 30$ for each product j. In Figure 3, Panels A and B, we plot the β_k coefficients from the volume and page views regressions and observe no significant irregularities in page views or review volume.

Controlling for Price, Promotions, and Product Description Length

Our data set lacks information on prices and product descriptions over time. To address this issue, we download historical prices and product descriptions from the Internet Archive, which is a nonprofit digital library that collects historical snapshots of web pages. We are able to find historical snapshots for 5,020 out of the 5,077 products in our data. In total we collect 145,564 snapshots of product pages, with an average of 29 snapshots per product. From each snapshot, we extract (1) the displayed price, (2) whether this price is marked down (based on the presence of the word "was" in the price field) and (3) the product description. We then associate each review in our sample with its chronologically nearest snapshot. We reestimate our main specifications with three additional time-varying controls: prices, whether there was a price promotion, and the length of the product description. We find that our estimates for the impact of Q&A, shown in Table 12, are robust to the inclusion of these controls.

Table 12. The Impact of Q&A on Ratings Controlling for Price, Discounts, and Product Description Lengths.

	IV (Second Stage)	IV (Second Stage)	IV (Second Stage)
POST × Low rating	.120***		
	(.041)		
POST × High variance		.084***	
		(.030)	
POST × Fit			1.111***
			(.352)
POST	−.014	−.028*	−.015
	(.012)	(.015)	(.013)
Review rank	.0001*	.0001*	.0001**
	(.0001)	(.0001)	(.0001)
On sale	−.025**	−.025**	−.025**
	(.011)	(.011)	(.011)
log(Price)	−.141***	−.140***	−.142***
	(.027)	(.027)	(.027)
log(Desc. length)	.026	.026	.025
	(.016)	(.017)	(.016)
Product FEs	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes
Observations	184,705	184,705	184,705

*$p < .1$.
**$p < .05$.
***$p < .01$.
Notes: FEs = fixed effects. Standard errors are clustered at the product level.

Changes in the Text of Product Descriptions

One concern with these estimates is that character counts are a crude summary of product descriptions. Product descriptions might remain the same length even though their content changes. Thus, we also investigate substantive changes in the contents of product descriptions within a 180-day window centered on each product's first answer. To do so, we turn each product description into a bag-of-words representation: a vector of word counts scaled by each word's inverse document frequency (a measure known as "tf-idf" in the natural language processing literature [Ramos 2003]). To quantify changes in product descriptions over time, we choose each product's first description as a reference point and compute cosine similarities between the first description and all subsequent descriptions. Finally, we estimate Equation 8 using these cosine similarities as our dependent variable. If product descriptions change following Q&A arrival, we would expect their cosine similarity to the reference description to decline. We report the results in Table 13. Overall, we see no decline in the similarity of product descriptions following the product's first answer.

Changes in Review Text

One might be concerned that fit uncertainty is resolved by consumer reviews that arrive alongside Q&As. This concern is partly mitigated by the fact that we control for the stock of pre-Q&A reviews received by each product using product fixed

Table 13. Cosine Similarity Around the First Answer.

	OLS (Rating Proxy)	IV (Second Stage)	OLS (Variance Proxy)	IV (Second Stage)	OLS (Fit Proxy)	IV (Second Stage)
POST × Low rating	.003	.018*				
	(.005)	(.009)				
POST × High variance			.008	.013		
			(.004)	(.007)		
POST × Fit					−.002	.083
					(.033)	(.067)
POST	−.003	−.006	−.006	−.007	−.003	−.004
	(.003)	(.004)	(.003)	(.004)	(.002)	(.004)
Product FEs	Yes	Yes	Yes			
Year-month FEs	Yes	Yes	Yes			
Observations	20,432	14,741	20,432	14,741	20,432	14,741

*$p < .05$.
Notes: FEs = fixed effects. Standard errors are clustered at the product level.

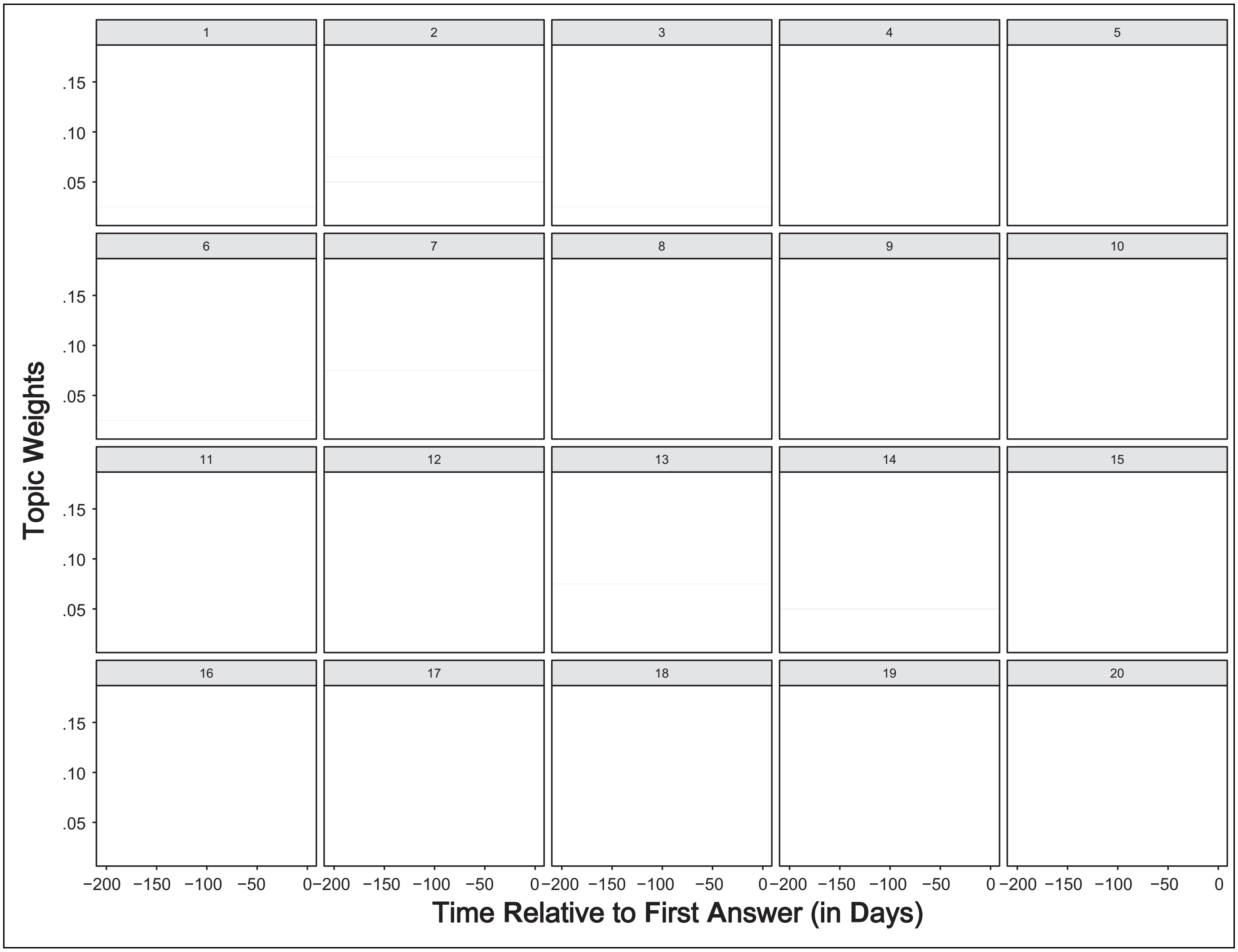

Figure 4. Plot of LDA topic weights for 20 topics, computed based on reviews from 0 to 200 days before Q&A.
Notes: We see no evidence of a change in review composition prior to Q&A arrival.

effects. However, an increased flow of fit-related reviews that coincides with the arrival of Q&As would bias our estimates. In this situation, we would expect to see a change in the composition of review text in the period leading to each product's first answer reflecting an increased focus of reviews on fit-related issues. To investigate changing trends in review text, we begin by fitting an LDA topic model with 20 topics on all reviews prior to Q&As (the topics are available in the Web Appendix).[16] We then group reviews on the basis of their arrival relative to each product's first answer. Finally, we calculate the average proportion of each topic within each group of reviews. We present these results in Figure 4. We observe that topic proportions remain relatively flat over time, leading us to believe that the contents of reviews do not abruptly change just before Q&A arrival.

Conclusion

In this article, we study how the Q&A technology of e-commerce platforms affects consumer choice. We start by providing an overview of ways in which Q&A and review corpora differ, highlighting the differential ability of Q&As to resolve fit uncertainty. Moreover, we show that negative reviews might often arise due to consumer–product fit mismatch. Our main finding is that answering consumer questions can lead to a subsequent increase in ratings, which is driven by a reduction in negative reviews that arise from fit mismatch.

Overall, our findings have direct managerial implications for platforms, retailers, and consumers. In the face of an ever-

[16] The choice of 20 topics is motivated by coherence score measures (Röder, Both, and Hinneburg 2015). We obtain qualitatively similar results with different topic numbers that yield similar coherence scores.

growing demand for information from online shoppers, it is important to realize potential positive synergies that might exist between different user-generated content features. Understanding these synergies can be important in guiding the platform's adoption decisions when designing different interactive elements to integrate into the product page. From the perspective of retailers, Q&As might be an effective communication tool that directly allows the retailer to interact with consumers before the purchase. In the case of management responses to reviews, which also offer an avenue of retailer–consumer communication, it is often not possible to remedy the "damage" done by negative reviews, because this communication occurs postpurchase. However, Q&As can serve as an effective reputation management tool from that perspective, because they can provide direct information that aids people in making better purchases and mitigates the risk of negative feedback. Finally, from the standpoint of consumers, Q&As can lead to better-informed purchases and, thus, higher postpurchase satisfaction, which can be expected to result in fewer product returns and more platform loyalty downstream.

Our results also have implications for the design of reputation systems. As we show, negative reviews can arise due to not only poor quality but also idiosyncratic fit mismatch. Platforms could consider running an algorithm similar to the classifier we propose, which could disambiguate these two broad classes of reviews, and compute a separate fit-based and quality-based average rating. This could be a way to mitigate the usual problems associated with biased average ratings on e-commerce platforms by leading customers to make more informed purchases.

A key limitation of our results is the inability to directly look at purchase behavior, because we do not have access to enough purchase level data. Some research has begun to examine the impact of Q&As on purchases and found evidence of a positive impact (Khern-am-nuai et al. 2020). Future research could investigate how Q&As affect different stages of the purchase funnel (from consideration to purchase) and the implications this has for firms and consumers. Another limitation is that we were unable to adopt a cross-platform identification strategy, which could have more robustly ruled out unobservable time-varying shocks. We were unable to find a comparable platform that sells the same products but does not have Q&A.

It is also important to note that Q&As are not costless—there is a possibility of information overload as more features accumulate on an e-commerce platform (e.g., videos and photos posted by users). Further, in our data set, most questions have a single answer, but this is starting to change: for certain platforms, almost all questions receive multiple answers. This might give rise to ambiguities and perhaps change the objective nature of the Q&A technology. Given these developments, it would be worth exploring the limits of the effect we observe and better understanding what constitutes too much information (Branco, Sun, and Villas-Boas 2015).

Appendix

Table A1. Top-Ten Product Categories Ordered by the Average Number of Questions Received.

Category	Questions per Product
Set top boxes	42.79
iPod	23.84
Telephones and accessories	23.20
Televisions and accessories	19.40
DVD players	14.67
Fitted kitchens	13.94
Large kitchen appliances	13.51
Satellite navigation and in-car entertainment	12.39
Heating and cooling	11.46
Kitchen electricals	10.83

Table A2. Review Volume Following the First Answer.

	OLS (Rating Proxy)	OLS (Variance Proxy)	OLS (Fit Proxy)
POST × Low rating	−.037		
	(.033)		
POST × High variance		.020	
		(.037)	
POST × Fit			−.157
			(.105)
POST	.089*	.074	.085*
	(.050)	(.054)	(.047)
Product FEs	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes
Weekday FEs	Yes	Yes	Yes
Observations	37,863	37,863	37,863

*$p < .1$.
Notes: FEs = fixed effects. Standard errors are clustered at the product level.

Table A3. Page View Volume Following the First Answer.

	OLS (Rating Proxy)	OLS (Variance Proxy)	OLS (Fit Proxy)
POST × Low rating	.318		
	(.468)		
POST × High variance		.607	
		(.479)	
POST × Fit			.710
			(.930)
POST	−.424	−.506	−.349
	(.410)	(.408)	(.349)
Product FEs	Yes	Yes	Yes
Year-month FEs	Yes	Yes	Yes
Weekday FEs	Yes	Yes	Yes
Observations	20,209	20,209	20,209

Notes: FEs = fixed effects. Standard errors are clustered at the product level.

Acknowledgments

The authors thank the Department of Marketing at Boston University; conference participants at CODE@MIT, WISE, ACM, EC, and EMAC; and Narendra Mukherjee for their valuable feedback. The authors also acknowledge the Wharton Customer Analytics Initiative for access to the data that made this research possible.

Associate Editor

Dina Mayzlin

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Georgios Zervas https://orcid.org/0000-0002-9558-244X

References

Acemoglu, Daron and Amy Finkelstein (2008), "Input and Technology Choices in Regulated Industries: Evidence from the Health Care Sector," *Journal of Political Economy*, 116 (5), 837–80.

Agichtein, Eugene, Yandong Liu, and Jiang Bian (2009), "Modeling Information-Seeker Satisfaction in Community Question Answering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3 (2), 1–27.

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis (2011), "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science*, 57 (8), 1485–1509.

Arellano, Manuel and Stephen Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58 (2), 277–97.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015), "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PloS One*, 10 (7), e0130140.

Benamara, Farah, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S. Subrahmanian (2007), "Sentiment Analysis: Adjectives and Adverbs Are Better Than Adjectives Alone," *ICWSM*, 7, 203–06.

Blei, David M, Andrew Y. Ng, and Michael I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3 (March), 993–1022.

Bondi, Tommaso (2019), "Alone, Together: Product Discovery Through Consumer Ratings," NET Institute Working Paper No. 19-09, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3468433.

Branco, Fernando, Monic Sun, and J. Miguel Villas-Boas (2015), "Too Much Information? Information Provision and Search Costs," *Marketing Science*, 35 (4), 605–18.

Chevalier, Judith A. and Dina Mayzlin (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43 (3), 345–54.

Cortes, Corinna and Vladimir Vapnik (1995), "Support-Vector Networks," *Machine Learning*, 20 (3), 273–97.

Dhanasobhon, Samita, Pei-Yu Chen, and Michael Smith (2007), "An Analysis of the Differential Impact of Reviews and Reviewers at Amazon.com," in *ICIS 2007 Proceedings*. Atlanta: Association for Information Systems.

Dimoka, Angelika, Yili Hong, and Paul A. Pavlou (2012), "On Product Uncertainty in Online Markets: Theory and Evidence," *MIS Quarterly*, 36 (2), 395–426.

Donald, Stephen G. and Kevin Lang (2007), "Inference with Difference-in-Differences and Other Panel Data," *Review of Economics and Statistics*, 89 (2), 221–33.

Gallino, Santiago and Antonio Moreno (2018), "The Value of Fit Information in Online Retail: Evidence from a Randomized Field Experiment," *Manufacturing & Service Operations Management*, 20 (4), 767–87.

Godes, David and José C. Silva (2012), "Sequential and Temporal Dynamics of Online Opinion," *Marketing Science*, 31 (3), 448–73.

Griliches, Zvi and Jerry A. Hausman (1986), "Errors in Variables in Panel Data," *Journal of Econometrics*, 31 (1), 93–118.

Gupta, Atul (2017), "Impacts of Performance Pay for Hospitals: The Readmissions Reduction Program," Becker Friedman Institute for Research in Economics Working Paper (October 16), https://bfi.uchicago.edu/working-paper/impacts-of-performance-pay-for-hospitals-the-readmissions-reduction-program/.

Hong, Yili and Paul A. Pavlou (2014), "Product Fit Uncertainty in Online Markets: Nature, Effects, and Antecedents," *Information Systems Research*, 25 (2), 328–44.

Hu, Nan, Paul A. Pavlou, and Jie Jennifer Zhang (2017), "On Self-Selection Biases in Online Product Reviews.," *MIS Quarterly*, 41 (2), 449–71.

Khern-am-nuai, Warut, Hossein Ghasemkhani, Dandan Qiao, and Karthik Natarajan Kannan (2020), "The Impact of Online Q&As on Product Sales: The Case of Amazon Answer," SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2794149.

Kwark, Young, Jianqing Chen, and Srinivasan Raghunathan (2014), "Online Product Reviews: Implications for Retailers and Competing Manufacturers," *Information Systems Research*, 25 (1), 93–110.

Lappas, Theodoros, Chrysanthos Dellarocas, and Neda Derakhshani (2017), "Reputation and Contribution in Online Question-Answering Communities," SSRN (February 16), https://ssrn.com/abstract=2918913.

Liu, Xiao, Dokyun Lee, and Kannan Srinivasan (2019), "Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning," *Journal of Marketing Research*, 56 (6), 918–43.

Luca, Michael (2016), "Reviews, Reputation, and Revenue: The Case of Yelp.com," HBR Working Paper 12-016, https://www.hbs.edu/ris/Publication%20Files/12-016_a7e4a5a2-03f9-490d-b093-8f951238dba2.pdf.

Manchanda, Puneet, Grant Packard, and Adithya Pattabhiramaiah (2015), "Social Dollars: The Economic Impact of Customer

Participation in a Firm-Sponsored Online Customer Community," *Marketing Science*, 34 (3), 367–87.

McAuley, Julian and Alex Yang (2016), "Addressing Complex and Subjective Product-Related Queries with Customer Reviews," in *International World Wide Web Conferences Steering Committee*. Switzerland: International World Wide Web Conferences Steering Committee, 625–35.

Narayanan, Sridhar and Harikesh S. Nair (2013), "Estimating Causal Installed-Base Effects: A Bias-Correction Approach," *Journal of Marketing Research*, 50 (1), 70–94.

Nickell, Stephen (1981), "Biases in Dynamic Models with Fixed Effects," *Econometrica: Journal of the Econometric Society*, 49 (6), 1417–26.

Proserpio, Davide and Georgios Zervas (2017), "Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews," *Marketing Science*, 36 (5), 645–65.

Ramos, Juan (2003), "Using tf-idf to Determine Word Relevance in Document Queries," *Proceedings of the First Instructional Conference on Machine Learning*, 242 (1), 29–48.

Ravi, Sujith, Bo Pang, Vibhor Rastogi, and Ravi Kumar (2014), "Great Question! Question Quality in Community Q&A," in *Proceedings of the International AAAI Conference on Web and Social Media*, Palo Alto, CA: AAAI.

Robinson, George K. (1991), "That BLUP Is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6 (1), 15–32.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015), "Exploring the Space of Topic Coherence Measures," in *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. New York: Association for Computing Machinery, 399–408.

Roodman, David (2009), "How to Do Xtabond2: An Introduction to Difference and System GMM in Stata," *Stata Journal*, 9 (1), 86–136.

Sahoo, Nachiketa, Chrysanthos Dellarocas, and Shuba Srinivasan (2018), "The Impact of Online Product Reviews on Product Returns," *Information Systems Research*, 29 (3), 723–38.

Senecal, Sylvain and Jacques Nantel (2004), "The Influence of Online Product Recommendations on Consumers' Online Choices," *Journal of Retailing*, 80 (2), 159–69.

Sun, Monic (2012), "How Does the Variance of Product Ratings Matter?" *Management Science*, 58 (4), 696–707.

Tirole, Jean (1988), *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

Zhu, Feng and Xiaoquan Zhang (2010), "Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics," *Journal of Marketing*, 74 (2), 133–48.