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

Shrabastee Banerjee*[1], Chrysanthos Dellarocas[1], and Georgios Zervas[1]

[1] *Boston University Questrom School of Business*

September 29, 2020

Abstract

We study 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, we show that Q&As complement consumer ratings and reviews: unlike reviews, questions are primarily asked pre-purchase, focus on clarification of product attributes rather than discussion of quality; answers convey fit-specific information in a predominantly sentiment-free way. Based on these observations, we hypothesize that Q&As mitigate product fit uncertainty, leading to better matches between products and consumers, and therefore improved product ratings. We find that when products suffering from fit mismatch start receiving Q&As, their subsequent ratings improve by approximately 0.1 to 0.5 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. Our 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: User Generated Content, Reputation Systems, E-Commerce, Q&A

*Corresponding author: sban@bu.edu

1 Introduction

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; Zhu and Zhang, 2010; Luca, 2016). 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 paper, 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 UK-based electronic commerce platform spanning a period of 5 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 upon, 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 from inadequate or wrong information on a retailer's website, highly individualized fit requirements on behalf of consumers, or intrinsic product complexity.

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 pre-purchase, 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

[1] Figure 6 in the Web Appendix show examples of Q&As from various platforms. Amazon displays Q&As prominently above consumer reviews on each product page.

be a complement to consumer reviews since it allows 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[2], 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 post-purchase, 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 discuss fit issues.

Using data from a major UK 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 0.1 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.

[2]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 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 dataset 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, which allows us to look at historical snapshots of the product pages in our sample. Based on this data, we re-estimate 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.

2 Related Work

The impact of ratings and reviews on consumer behavior (most notably, purchase decisions) has been well-documented in the literature. For example, in an online experiment, it was shown 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 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 paper we do not directly look at purchases, these studies show that an increase in average ratings is a positive and managerially relevant outcome, since it has been widely shown to correlate 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 et al., 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 looked at. Most of the work in the domain of Q&As has focused on question-answering communities, such as Quora and StackOverflow. Questions examined in this area include: how reputation relates to response volume, question difficulty and answer quality on Stack Overflow (Lappas et al., 2017), how to model the satisfaction of information seekers in Q&A communities (Agichtein et al., 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, it has been 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 et al., 2015). In such a setting, the source of economic benefit is seen 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 paper, 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.com (McAuley and Yang, 2016). However, the focus of this work is developing and comparing the algorithm to other existing text mining tools, and does not investigate any causal questions that combine reviews and Q&As.

Our conceptual framework 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 et al., 2012). Products may have inherent quality issues that are revealed only through prolonged product usage - hence, reviews can be a valuable avenue through which quality uncertainty is mitigated.

Product fit uncertainty, on the other hand, arises because buyers cannot easily assess whether the product's characteristics match their requirements or tastes (Hong and Pavlou,

2014; Kwark et al., 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. Hence, 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, it was seen 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: it has been 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 et al., 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, it has been shown that reviews on Goodreads.com 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, since there are fewer objective attributes, and open-ended reviews might be more helpful in terms of resolving fit uncertainty (Bondi, 2019).

3 Conceptual Framework

Uncertainty in the context of e-commerce can be thought to be 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, we hypothesize that reviews are not as well-suited to reducing fit uncertainty because 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 was 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 her idiosyncratic fit uncertainty prior to purchase. In cases such as the above, we hypothesize that 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 post-purchase regrets among those who choose to purchase.

In the mathematical appendix (Section A.2) we present a model that captures how the presence of informative Q&A affects consumer decision making and product ratings in settings with consumer fit uncertainty. For the sake of simplicity, we are abstracting away quality uncertainty so we can more cleanly focus on the effects of Q&A on reducing fit uncertainty. Adding quality uncertainty to the model will result in a presence of some quality-related negative ratings both without and with Q&A (and this, of course, is what we observe in our data). However, all insights about the increase in average ratings with Q&A, due to the reduction in fit-related negative ratings, will remain the same. Below we summarize the model's assumptions and key predictions.

We model fit uncertainty by assuming that, for every product, a fraction of consumers are uninformed about one or more product attributes that are of idiosyncratic value to them. Each uninformed consumer may care about different attributes and may value their states differently. For example, consider the case of external hard drives and assume that compatibility of a hard drive with a computer architecture is the unknown attribute. For PC

users the desirable state of this attribute would be PC compatible whereas for Mac users the desirable state would be Mac compatible. In the absence of additional information, uninformed consumers form prior beliefs about the fit uncertainty (denoted in the mathematical model by α). These beliefs may reflect the consumers' broad understanding of the distribution of product attributes within a product category. For example, if it is known that 70% of external hard drives on the market are PC-compatible and 30% are Mac-compatible, uninformed PC users would be justified in having a 30% prior expectation of fit mismatch, whereas uninformed Mac users would have a corresponding 70% fit mismatch expectation. We assume that purchasers who experience good fit walk away with positive utility and leave positive ratings whereas those who experience poor fit walk away with negative utility and leave negative ratings.[3] Our model shows that without Q&As, the behavior of uninformed consumers can fall into either an *optimistic* or a *pessimistic* case. In the former, uninformed consumers choose to purchase because their prior beliefs about fit mismatch are relatively optimistic (low α) and/or the consequences of fit mismatch are relatively mild (for example, the product can be returned easily). With some probability, these consumers experience fit mismatch and leave negative feedback. On the other hand, for the latter (pessimistic) case, uninformed consumers choose not to purchase, because their prior beliefs are not optimistic (high α) and/or the consequences of fit mismatch are severe (for example, returning the product is not easy or fit mismatch results in damage to property or health). In this case, only informed consumers purchase leaving positive feedback.

The presence of Q&A allows uninformed consumers to request and receive information about unknown product attributes. Our model shows that, in order for Q&A to change consumer behavior, the reliability of answers needs to be sufficiently high. Under that condition, Q&A affects consumer behavior in two ways: it prevents optimistic consumers from purchasing products that would have been a bad fit for them, or it encourages pessimistic consumers to purchase products that would be a good fit for them.

The first effect is responsible for the increase in product ratings.[4] In the limiting case where Q&A are perfectly informative (that is, where answers are always correct) our model

[3]The model can be extended to a multi-valued rating scale $1, 2, ..., n$ by defining a correspondence between post-purchase utilities $u_1, u_2, ..., u_{n-1}$, where $u_i < u_{i+1}$, such that consumers post rating i if they experience post-purchase utility $u_{i-1} < u \leq u_i$ plus the obvious corner cases. The precise thresholds u_i may differ among consumers. Such a mapping retains the key properties that drive our stylized model, i.e. average ratings are positively related to average post-purchase utility and negatively related to the probability of fit mismatch among purchasers.

[4]Our model allows for both optimistic and pessimistic uninformed consumers to co-exist and predicts an increase in average ratings as long as the fraction of consumers belonging to the optimistic case (γ in the model) is greater than 0, and answers are reliably correct. In the external hard drive example we discussed above, PC users have reasons to be optimistic, whereas Mac users have reasons to be pessimistic. Both segments would typically co-exist in the consumer population for external hard drives.

predicts that the presence of Q&A completely eliminates fit mismatch and that this would result in perfectly positive ratings. In that case, 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, which, in turn, is proportional to the amount of fit-related negative reviews without Q&A. Our theory predicts that fit-related negative reviews are likely to be higher when:

1. there are many uninformed consumers, that is, the product exhibits a higher fit uncertainty, and

2. the product is a bad fit for a large fraction of uninformed consumers, that is, the product caters to niche tastes that do not coincide with the mainstream, and

3. many uninformed consumers are optimistic about the probability of a good fit and choose to purchase in the presence of fit uncertainty; as previously discussed, this happens when most products of this category are a good fit for most consumers and/or when then impact of bad fit is not very severe relative to the utility of a good fit.

Continuing our running example, external hard drives that are only compatible with Mac computers illustrate the above conditions. Assuming that 1) a hard drive's compatibility is not clearly specified in the product description, 2) the majority of prospective buyers are PC users, and 3) most external hard drives on the market are compatible with PCs, Mac-only hard drives constitute an example of a product that exhibits bad fit for most uninformed consumers in a product category where the majority of uninformed consumers (the PC users) have rational reasons to feel optimistic enough to purchase. In the presence of fit uncertainty, Mac-only hard drives will accumulate a lot of negative fit-related reviews and stand to benefit a lot from compatibility clarifications made possible through Q&A.

These results hold approximately if the reliability of answers is imperfect, but high. This is the assumption we will adopt in the rest of the paper.

4 Data and descriptives

We obtain data from Bazaarvoice via the Wharton Consumer Analytics Initiative[5]. Bazaarvoice provides software that enables businesses to collect and display reviews and Q&As on their websites. Our data comes from a UK-based big-box retailer (similar to Amazon.com) that uses Bazaarvoice software. The data covers two product categories (Technology and Home & Garden) and includes all reviews and Q&As posted between 2009 and 2015.

[5]https://wcai.wharton.upenn.edu/

The two product categories are further subdivided into 755 subcategories such as Bedroom Furniture and Video Games.

Overall, our data contains 37,853 unique product identifiers.[6] 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. Since we want to study the impact of Q&As on product ratings, our analyses will 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 to the above, we make use of click-stream data collected over a two month period in 2015 (February and March) to supplement our main analyses. This data consists of the browsing behavior of customers (i.e. which specific product pages were clicked on). Within this dataset, we look at products that received their first answer within the two month observation period.

We also collect data from the Internet Archive to conduct a series of robustness checks. These supplementary data sets are described in detail in Section 7.

4.1 Q&A, reviews, and fit uncertainty

The hypotheses we develop in this paper relate to the ability of Q&As to resolve fit uncertainty. In this section, we show that the Q&A corpus has a number of characteristics that make it particularly well-suited to conveying information about fit to consumers. We also compare Q&As to 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, since the ability of Q&As to resolve fit uncertainty depends on the rate at which the feature is used by consumers. Figure 1 provides some descriptive patterns of Q&A dynamics. In Figure 1a and Figure 1b, we find evidence of increased usage of Q&As over time, mirroring the increasing adoption of reviews. In Figure 1c, 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 gone

[6]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. Additionally, allowing for a product-level rather than a product-group level fixed effect in our estimation lends more flexibility to the model.

down from 17 days in 2011 to 4 days in 2015, suggesting increasing engagement with the feature. Finally, in Figure 1d, 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.[7]

In Table 2, 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. Since 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 pre-purchase, fit-specific information, and do so in a less sentiment-laden fashion than reviews. First, to get at the pre-purchase 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 10 questions each. We asked coders whether the question was most likely asked before or after purchase.[8] 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 pre-purchase. This is in stark contrast with reviews, which occur almost always (and for some platforms, exclusively) post-purchase.

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 to 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 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, while 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

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

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

do this, we use a Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003).[9] 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 above, the topics obtained for Q&As contain more references to fit-related attributes (such as 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 3.[10]

5 Empirical Strategy

In this section, we begin by motivating our main estimating equation, and then discuss our identification strategy for estimating the causal impact of Q&As on consumer reviews. We assume that products are both vertically and horizontally differentiated. Thus, each purchaser i derives post-purchase 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 purchaser i's preference. Thus, ex-post utility takes the form:

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

Our main hypothesis, articulated in Section 3, is that Q&As can provide fit-specific information that allows purchasers to buy products better suited to their idiosyncratic needs, resulting in higher post-purchase utility.

To capture the potential impact of Q&As, we model the horizontal component of post-purchase utility as:

$$\mu_{ij} = \beta_j \cdot \text{POST}_{ij} + \epsilon_{ij}, \tag{2}$$

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:

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

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

[10] For the complete set of topics see Table 15 and Table 16 in appendix A

where m_j captures the *reduction* to the degree of fit mismatch faced by purchasers of product j due to Q&A. 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. As described in Section 3 and Section A.2, m_j is proportional to the probability of fit mismatch *without* Q&A .[11]

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:

$$U_{ij} = \eta_j + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (m_j \cdot \text{POST}_{ij}) + \epsilon_{ij}. \tag{4}$$

In our data, we observe individual reviews left by purchasers of each product j rather than utility. Thus, under the assumption that ratings r_{ij} are an increasing function of utility, we modify Equation 4 above to arrive at the following estimating equation:

$$r_{ij} = \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (m_j \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \epsilon_{ij}. \tag{5}$$

Compared to the utility function above, this equation 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 Section 7, we discuss specific threats to this assumption and perform robustness checks.

5.1 Measuring fit mismatch

Our theoretical framework predicts that the effect of Q&A on product ratings is proportional to the probability that purchasers will experience fit mismatch without Q&A, which, in turn, is proportional to the amount of fit-related negative reviews without Q&A. However, the degree of fit mismatch inherent in any product's past purchases is not directly observable, since we do not know which negative reviews arise due to poor quality, fit mismatch,

[11] As Section A.2 shows, the post-Q&A increase in average purchaser utility (and in turn ratings) is due to the post-Q&A reduction in mismatch probability among purchasers. This, in turn, is proportional to the pre-Q&A mismatch probability as well as the amount of fit related negative reviews.

or other reasons. We approach this problem by constructing three proxies for the presence of fit mismatch prior to Q&A, which we describe below.

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. Hence, 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 rating of 4 out of 5 as the threshold that separates these products that may suffer from fit mismatch from those that don't. Later, 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 (for example, low quality products). These false positives may attenuate the Average Treatment effect on the Treated (ATT) we estimate.

High rating variance Our second measure looks at products that have a high rating variance (Sun, 2012). We can think of such products as suffering from fit mismatch since they have attributes that are asymmetrically preferred by consumers (some like them and find them to be a good fit, others don't). 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. Based 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 most the likely cause of 3,300 randomly chosen negative (1-, 2-, and 3-star) reviews.[12] We indicate three categories into which reviews are to be classified: poor fit, poor quality, or other miscellaneous reasons (such as issues with the store or shipping). Figure 7 in the Web Appendix displays the survey seen by the coders. Any disagreements are 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. Since 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 labelled reviews to train a C-Support Vector Machine (C-SVM) classifier (Cortes and Vapnik, 1995). To perform the classification task, we remove common stopwords, and then tokenize and stem the text of each negative review into a bag-of-words representation, thus 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.[13]

We train our classifier on an 80% random sample of our labelled 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 having a larger-margin separating hyperplane. We select a value for C using 5-fold grid search cross-validation. We evaluate the out-of-sample performance of our classifier using the commonly employed ROC-AUC (receiver operating characteristic area under the curve) metric. ROC-AUC ranges from 0 to 1 and it is a ranking metric. Intuitively, a 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 a ROC-AUC of 0.82.[14]

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 4 we present the top-5 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

[12] Refer to Table 19 in the Web Appendix for some illustrative examples of reviews arising from poor quality vs poor fit.

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

[14] We also replicate our main analyses with a Naive Bayes classifier, which achieves an ROC-AUC of 0.79. These results are available upon request.

requirements that the product failed to fulfill, despite not inherently being of an inferior quality.

Second, we order all product categories in our dataset 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 hard 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. On the other hand, products where the customer's domain knowledge dictates their purchase (such as DIY and power tools) have fewer fit concerns according to our classifier.

Third, we examine the top 20 words that are most predictive of fit vs. 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 SVM's, the LRP score of each word-class combination is computed as the sum of the products of the word loading and the tf-idf 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 the classes in Table 5. We find, as expected, that words which 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.

5.2 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., $\mathbb{E}[\text{POST}_{ij} \cdot \epsilon_{ij}] = 0$. Here, we discuss these two problems under the assumption of treatment exogeneity; we discuss violations to treatment exogeneity separately in Section 7.

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 $m_j \cdot \text{POST}_{ij}$, violating strict exogeneity and biasing OLS estimates of β_1. (To see this, note that demeaning $m_j \cdot \text{POST}_{ij}$ 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 directly, instead relying on noisy measures $\tilde{m}_j = m_j + v_j$ (where $\tilde{m}_j$ 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 $\tilde{m}_j$, we have:

$$\begin{aligned} r_{ij} &= \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (m_j \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \epsilon_{ij} \\ &= \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot ((\tilde{m}_j - v_j) \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \epsilon_{ij} \\ &= \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot (\tilde{m}_j \cdot \text{POST}_{ij}) + X'_{ij} \cdot \beta_2 + \tilde{\epsilon}_{ij}, \end{aligned} \tag{6}$$

where

$$\tilde{\epsilon}_{ij} = -\beta_1 \cdot (v_j \cdot \text{POST}_{ij}) + \epsilon_{ij}. \tag{7}$$

Note that $\text{Cov}(\tilde{m}_j \cdot \text{POST}_{ij}, \tilde{\epsilon}_{ij}) \neq 0$, since 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 instrument.[15] Specifically, we divide the pre-Q&A period for each product into two smaller samples: a hold-out period, which includes all reviews up to 200 days prior to the first answer, and a shorter pre-treatment period that includes all reviews starting at 200 days prior to the first answer and ending at the time of the first answer.[16] We then use these two samples to construct the two proxies $\tilde{m}_j^{\text{hold-out}}$ and $\tilde{m}_j^{\text{pre}}$, where the former quantity can be thought of as a lag of the latter. Finally, we use $\tilde{m}_j^{\text{hold-out}}$ to construct instruments for

[15] For examples of recent empirical work that has used similar strategies see Acemoglu and Finkelstein (2008), and Gupta (2017)

[16] In a robustness check, we change the definition of our hold-out sample to make it more flexible: out of all pre-Q&A observations, we select the most recent 50% to form the pre-treatment sample, and the rest to form the hold-out sample. We report these estimates, which are similar to our main results in Table 20 - Table 22 of the Web Appendix.

the endogenous variable $\tilde{m}_j^{\text{pre}} \cdot \text{POST}_{ij}$. The holdout sample is subsequently excluded from estimation. Our main estimating equation and the corresponding first stage are given by:

$$r_{ij} = \eta_j + \delta_{t(ij)} + \beta_0 \cdot \text{POST}_{ij} + \beta_1 \cdot \widehat{(\tilde{m}_j^{\text{pre}} \cdot \text{POST}_{ij})} + X'_{ij} \cdot \beta_2 + \tilde{\epsilon}_{ij}, \tag{8}$$

$$(\tilde{m}_j^{\text{pre}} \cdot \text{POST}_{ij}) = \tilde{\eta}_j + \tilde{\delta}_{t(ij)} + \gamma_0 \cdot \text{POST}_{ij} + \gamma_1 \cdot (\tilde{m}_j^{\text{hold-out}} \cdot \text{POST}_{ij}) + X'_{ij} \cdot \gamma_2 + \tilde{u}_{ij}. \tag{9}$$

where $\tilde{\epsilon}_{ij} = -\beta_1 \cdot (v_j^{\text{pre}} \cdot \text{POST}_{ij}) + \epsilon_{ij}$ and $\tilde{u}_{ij} = -\gamma_1 \cdot (v_j^{\text{hold-out}} \cdot \text{POST}_{ij}) + u_{ij}$. To see why this strategy works, notice that:

$$\tilde{m}_j^{\text{hold-out}} = m_j + v_j^{\text{hold-out}}, \tag{10}$$

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

The instrument $\tilde{m}_j^{\text{hold-out}} \cdot \text{POST}_{ij}$ 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 $\mathbb{E}[(\tilde{m}_j^{\text{hold-out}} \cdot \text{POST}_{ij})\tilde{\epsilon}_{ij}] = 0$. This condition will be met as long as the two measurement errors are not correlated, i.e., $\mathbb{E}[v_j^{\text{hold-out}} \cdot v_j^{\text{pre}}] = 0$ (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 pre-treatment period were not more likely to also experience such shocks in the hold-out period. Thus, by instrumenting with hold-out ratings we use the signal embedded in hold-out measures of fit mismatch (m_j) to get rid of the noise in pre-treatment measures of fit mismatch (v_j^{pre}), the latter being the source of bias when we estimate Equation 6 by OLS.

While 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 (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 $v_j^{\text{hold-out}}$ and v_j^{pre}) are not correlated.

Using BLUP to construct instruments As we discussed above, the proxies we use for fit mismatch are measured with error. Beyond causing issues with identification, measurement error means that the instrument $\tilde{m}_j^{\text{hold-out}} \cdot \text{POST}_{ij}$ may be a poor predictor of $\tilde{m}_j^{\text{pre}} \cdot \text{POST}_{ij}$ for products with few reviews in the hold-out 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 $\tilde{m}_j^{\text{hold-out}}$, where we shrink the estimates of $\tilde{m}_j^{\text{hold-out}}$ for products with few reviews towards the population mean (Robinson, 1991). Specifically, for each product we estimate the best linear unbiased predictor (BLUP) of its mean rating in the hold-out sample using a mixed effects model with a random intercept m_j for each product j, and a fixed intercept μ:

$$\tilde{m}_j^{\text{hold-out}} = \mu + m_j + e_j$$

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 hold-out sample for each product, which we use to construct our final instruments.[17]

6 Results

Now, we turn to estimating the effect of Q&A arrival on ratings using each of the three fit mismatch measures that we described above.

Low pre-Q&A ratings Our first measure is low pre-treatment average ratings. Hence, we estimate Equation 8 with $\tilde{m}_j^{\text{pre}}$ being an indicator for products with low average pre-treatment ratings (≤ 4), and $\tilde{m}_j^{\text{hold-out}}$ being the average rating of the product in the hold-out period. In addition to product and time fixed effects, we also control for the rank of each review (as recommended for example by Godes and Silva (2012)).

We present our results in Table 6. Column 1, which presents our OLS estimates, serves as the baseline and includes all products and all reviews in the estimation sample. We find a positive and significant increase in ratings of 0.24 for products with a low pre-treatment mean. We also find a statistically significant decrease of 0.045 stars for products that have a high pre-treatment mean. As discussed in Section 5.2, some products may have high or low pre-treatment ratings by pure chance rather than as a consequence of poor fit. We would expect the ratings of these product to mean revert regardless of Q&As, which would inflate our estimates. In column 2, we re-estimate the OLS model by excluding the hold-out 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 IV specification

[17] 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.

described in Section 5.2. Column 3 reports the first stage of Equation 8. We see that average ratings computed based on the hold-out sample using BLUP are strong predictors of the pre-treatment average rating. Column 4 reports our preferred IV estimate: we find a positive and significant increase of 0.12 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 (hence we do not include the main effect for POST_{ij}). We instrument each of these variables with the corresponding lagged version from the hold-out sample. We find substantial heterogeneity: the higher the fit mismatch, the larger a product's post-treatment increase in ratings.

To put the magnitude of our effect—approximately 0.1 stars on average—in context, we compare it against the standard deviation of average ratings, which, for products with at least 5 reviews, is also roughly equal to 0.1 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 $\tilde{m}_j^{\text{pre}}$ being an indicator for products with pre-treatment variance ≥ 1 (the median value) and $\tilde{m}_j^{\text{hold-out}}$ being the rating variance in the hold-out period. We report our results in Table 7 for the full sample OLS (column 1), OLS excluding hold-out 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.

As described in Section 5.1, 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. Specifically, we estimate a specification that contains the full interaction between the rating and variance proxies resulting in four groups of products based on their pre-treatment 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 hold-out equivalent. Given our theory, we expect low rating/high variance products to be primarily impacted by Q&As. We present our results in Table 8. Among the four groups of products, we see a statistically significant increase in ratings following Q&A arrival only for the low rating/high variance group, as expected.

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 Section 5.1. We estimate Equation 8 with $\tilde{m}_j^{\text{pre}}$ being the fraction of pre-treatment fit related negative reviews and $\tilde{m}_j^{\text{hold-out}}$ being the fraction of fit related negative reviews constructed in the hold-out sample estimated using a logistic BLUP. We present our results in Table 9. 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 pre-treatment reviews expressing fit concerns, we estimate a subsequent increase in ratings of $1.135 \times 0.1 = 0.11$ stars.

6.1 Mechanism: fit mismatch reduction

We now turn to examining 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 Section 5.1). We code all non-fit-related negative reviews and all positive reviews (4 and 5 stars) as 0. 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).

As before, we present results for both OLS and IV specifications in Table 10, 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 estimates as follows: if 10% of all pre-treatment reviews are due to fit uncertainty, the product would experience a subsequent decline in the probability of receiving a negative fit-related review of -0.19 times 0.1, 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 (1-, 2-, and 3-stars.) However, conditional on receiving a negative review post-Q&A, the probability that this review is fit-related decreases by $\frac{100}{15} \times 1.9 = 12.6\%$.

Here we have shown that Q&As lead to fewer negative reviews that contain fit-related concerns. Our hypothesized theory for this reduction (as described in Section 3) 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 prior to 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 11.)

7 Robustness checks

Our results above 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 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 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 pageviews around each product's first answer, which we would expect in the presence of increased advertising. Next, we collect additional data that allows 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 dataset, we also show that the

content of product descriptions doesn't 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 below.

Changes in review volume or pageviews 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 been taking place around the time of each products' first answer.[18] We focus on two outcomes suggestive of increased marketing activity: the daily number of reviews, and the daily number of pageviews each products 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 11. We find no significant change in review volume around the first answer.

One concern with the analysis above is low power. Because reviews are relatively infrequent, a change in review volume can be difficult to detect. To increase power, we use click-stream data made available to us for a two month period (February and March 2015), and repeat our analysis using daily pageviews — 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 12. We see no significant change in pageviews post treatment.[19]

Finally, we graphically examine any changes in review volume or pageviews in the immediate neighborhood of the first answer. To do so, we estimate the following model for our two discrete measures of fit mismatch, low rating and high variance:

$$y_{jt} = \eta_j + \delta_t + \sum_{k=-30}^{30} \beta_k \times \mathbf{I}\{D_{jt} = k\}_j + \epsilon_{jt} \tag{12}$$

where y_{jt} is respectively the number of reviews or pageviews and $\mathbf{I}\{D_{jt} = k\}$ is an indicator for day $k \in -30, 30$ for each product j. In Figure 3 and Figure 4, we plot the β_k coefficients from the volume and pageviews regressions, and observe no significant irregularities

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

[19] We further address the issue of low power by using OLS on the full sample (results reported in Table 17 and Table 18 of appendix A), 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.

in pageviews or review volume.

Controlling for price, promotions, and product description length Our dataset lacks information on prices, and product descriptions over time. To deal with this issue, we download historical prices and product descriptions from the Internet Archive (IA), which is a non-profit 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 (a) the displayed price, (b) whether this price is marked down (based on the presence of the word "was" in the price field) and (c) the product description. We then associate each review in our sample with its chronologically nearest snapshot. We re-estimate 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 13, are robust to the inclusion of these controls.

Changes in the text of product descriptions One concern with the above estimates is that character counts are a crude summary of product descriptions. Product description 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 14. Overall, we see no decline in the similarity of product descriptions following the each 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 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 Table 23 in the Web Appendix)[20] We then group reviews based on their arrival relative to each 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 5. We observe that topic proportions remain relatively flat over time, leading us to believe that the contents of reviews do not abruptly change just prior to Q&A arrival.

8 Conclusion

In this paper, 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 as well as consumers. In the face of an ever-growing demand for information from online shoppers, it is important to realize potential positive synergies that might exist between different UGC 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 them to interact with consumers before purchases happen. 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, since this communication is post-purchase. However, Q&As can serve as an effective reputation management tool from that perspective, since they can serve to provide direct information that aids better purchases and mitigates the risk of negative feedback. Finally, from the standpoint of consumers, Q&As can lead to better informed purchases and hence higher post-purchase 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 not just due to poor quality, but due to idiosyncratic fit mismatch.

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

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 (Amazon.com has already started doing this). This could be a possible 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, since we do not have access to enough purchase level data. Some research has started to examine the impact of Q&As on purchases, and found evidence of a positive impact (Khern-am nuai et al., 2017). 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 not being able to adopt a cross-platform identification strategy, which would have been able to more robustly rule 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 exists the possibility of potential information overload as more features accumulate on an e-commerce platform (such as videos and photos posted by users). Further, in our dataset, 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/direct 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 et al., 2015).

On the whole, UGC implementations in general, and Q&A technology in particular, pose interesting problems for e-commerce websites that are worthy of further exploration.

References

Acemoglu, D. and Finkelstein, A. (2008). Input and technology choices in regulated industries: Evidence from the health care sector. *Journal of Political Economy*, 116(5):837–880.

Agichtein, E., Liu, Y., and Bian, J. (2009). Modeling information-seeker satisfaction in community question answering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(2):10.

Archak, N., Ghose, A., and Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8):1485–1509.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., and Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*. Citeseer.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bondi, T. (2019). Alone, together: Product discovery through consumer ratings. *Available at SSRN 3468433*.

Branco, F., Sun, M., and Villas-Boas, J. M. (2015). Too much information? information provision and search costs. *Marketing Science*, 35(4):605–618.

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dhanasobhon, S., Chen, P.-Y., Smith, M., and Chen, P.-y. (2007). An analysis of the differential impact of reviews and reviewers at amazon. com. *ICIS 2007 Proceedings*, page 94.

Dimoka, A., Hong, Y., and Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence. *MIS quarterly*, 36.

Donald, S. G. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, 89(2):221–233.

Gallino, S. and Moreno, A. (2018). The value of fit information in online retail: Evidence from a randomized field experiment. *Manufacturing & Service Operations Management*, 20(4):767–787.

Godes, D. and Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473.

Griliches, Z. and Hausman, J. A. (1986). Errors in variables in panel data. *Journal of econometrics*, 31(1):93–118.

Gupta, A. (2017). Impacts of performance pay for hospitals: The readmissions reduction program. *Becker Friedman Institute for Research in Economics Working Paper*, (2017-07).

Hong, Y. and Pavlou, P. A. (2014). Product fit uncertainty in online markets: nature, effects, and antecedents. *Information Systems Research*, 25(2):328–344.

Khern-am nuai, W., Ghasemkhani, H., and Kannan, K. N. (2017). The impact of online q&as on product sales: The case of amazon answer. *Available at SSRN 2794149*.

Kwark, Y., Chen, J., and Raghunathan, S. (2014). Online product reviews: Implications for retailers and competing manufacturers. *Information systems research*, 25(1):93–110.

Lappas, T., Dellarocas, C., and Derakhshani, N. (2017). Reputation and contribution in online question-answering communities. *Available at SSRN: https://ssrn.com/abstract=2918913*.

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp. com. *HBR Working Paper 12-016*.

Manchanda, P., Packard, G., and Pattabhiramaiah, A. (2015). Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Marketing Science*, 34(3):367–387.

McAuley, J. and Yang, A. (2016). Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of the 25th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, pages 1417–1426.

Proserpio, D. and Zervas, G. (2017). Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science*, 36(5):645–665.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

Ravi, S., Pang, B., Rastogi, V., and Kumar, R. (2014). Great question! question quality in community q&a. In *ICWSM*.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical science*, pages 15–32.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Roodman, D. (2009). How to do xtabond2: An introduction to difference and system gmm in stata. *The stata journal*, 9(1):86–136.

Sahoo, N., Dellarocas, C., and Srinivasan, S. (2018). The impact of online product reviews on product returns. *Information Systems Research*, 29(3):723–738.

Senecal, S. and Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of retailing*, 80(2):159–169.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, 58(4):696–707.

Tirole, J. (1988). *The Theory of Industrial Organization*. MIT Press.

Zhu, F. and Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, 74(2):133–148.

Tables

Table 1: Summary statistics.

	Products	Avg. rating	Reviews	Questions
Technology				
Products with both Q&A and reviews	1175	4.37 (0.55)	153.6 (138.5)	23.78 (34.84)
Home & Garden				
Products with both Q&A and reviews	3902	4.28 (0.436)	264.5 (242.9)	14.2 (16.3)

Table 2: Top-10 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
Sat-nav and in-car entertainment	12.39
Heating and cooling	11.46
Kitchen electricals	10.83

Table 3: Top-3 LDA topics for reviews and Q&As.

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

Table 4: Top-5 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

Table 5: Top-20 words most predictive of quality/fit issues, estimated using layerwise 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

Table 6: The impact of Q&A on ratings using low pre-treatment ratings (≤ 4) as a measure for fit mismatch.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1^{st} stage)	IV (2^{nd} stage)	IV (2^{nd} stage, bins)
POST × Low Rating	0.243^{***}	0.206^{***}		0.122^{***}	
	(0.022)	(0.031)		(0.041)	
POST × Hold-out Rating			-0.901^{***}		
			(0.029)		
POST	-0.045^{***}	-0.026^{**}	4.049^{***}	-0.011	
	(0.009)	(0.012)	(0.130)	(0.012)	
POST × Rating $\in [2, 3]$					0.502^{***}
					(0.101)
POST × Rating $\in (3, 4]$					0.100^{***}
					(0.037)
POST × Rating $\in (4, 5]$					-0.012
					(0.012)
Review Rank	0.0002^{***}	0.0001^{*}	-0.0001^{*}	0.0001^{*}	0.0001^{*}
	(0.0001)	(0.0001)	(0.00004)	(0.0001)	(0.0001)
Product FE	Yes	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes	Yes
F Statistic			312.91		
Observations	345,168	184,811	184,811	184,811	184,811

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 7: The impact of Q&A on ratings using high pre-treatment rating variance (≥ 1) as a measure for fit mismatch.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1^{st} stage)	IV (2^{nd} stage)
POST × High Variance	0.159***	0.150***		0.085***
	(0.015)	(0.019)		(0.030)
POST × Hold-out Variance			0.533***	
			(0.016)	
POST	−0.061***	−0.053***	−0.064***	−0.026*
	(0.010)	(0.012)	(0.019)	(0.015)
Review Rank	0.0001***	0.0001	0.00002	0.0001*
	(0.00004)	(0.0001)	(0.00004)	(0.0001)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			380.08	
Observations	345,168	184,811	184,811	184,811
Note:				*p<0.1; **p<0.05; ***p<0.01
				Standard errors clustered at the product level.

Table 8: The impact of Q&A on ratings using both low pre-treatment ratings (≤ 4) and high pre-treatment rating variance (≥ 1) as measures for fit mismatch.

	IV (2^{nd} stage)
POST × Low Rating × Low Variance	−0.594
	(0.429)
POST × Low Rating × High Variance	0.125***
	(0.038)
POST × High Rating × Low Variance	−0.010
	(0.014)
POST × High Rating × High Variance	−0.013
	(0.028)
Review Rank	0.0001*
	(0.0001)
Product FE	Yes
Year-month FE	Yes
Observations	184,811
Note:	*p<0.1; **p<0.05; ***p<0.01
	Standard errors clustered at the product level.

Table 9: The impact of Q&A on ratings using the pre-treatment fraction of reviews mentioning fit issues as a measure for fit mismatch.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1^{st} stage)	IV (2^{nd} stage)
POST × Fit	1.932***	1.948***		1.135***
	(0.150)	(0.288)		(0.356)
POST × Hold-out Fit			1.452***	
			(0.054)	
POST	−0.039***	−0.030**	−0.007***	−0.013
	(0.010)	(0.012)	(0.001)	(0.013)
Review Rank	0.0002***	0.0001**	−0.00000**	0.0001**
	(0.00005)	(0.0001)	(0.00000)	(0.0001)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			239.46	
Observations	345,168	184,811	184,811	184,811

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 10: Mechanism: products with a high fraction of pre-treatment fit related negative reviews experience a decline in such reviews following Q&A.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1^{st} stage)	IV (2^{nd} stage)
POST × Fit	−0.726***	−0.590***		−0.190***
	(0.024)	(0.065)		(0.072)
POST × Hold-out Fit			1.452***	
			(0.054)	
POST	0.016***	0.013***	−0.007***	0.005***
	(0.001)	(0.001)	(0.001)	(0.002)
Review Rank	−0.00002***	−0.00001	−0.00000**	−0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			239.46	
Observations	345,168	184,811	184,811	184,811

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 11: Review volume around the first answer.

	IV (2^{nd} stage)	IV (2^{nd} stage)	IV (2^{nd} stage)
POST × Low Rating	−0.033		
	(0.071)		
POST × High Variance		0.020	
		(0.052)	
POST × Fit			−0.543
			(0.732)
POST	0.094	0.080	0.100
	(0.075)	(0.078)	(0.076)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes
Observations	25,257	25,257	25,257
Note:			*p<0.1; **p<0.05; ***p<0.01
	Standard errors clustered at the product level.		

Table 12: Pageviews around the first answer.

	IV (2^{nd} stage)	IV (2^{nd} stage)	IV (2^{nd} stage)
POST × Low Rating	3.708		
	(4.191)		
POST × High Variance		1.413	
		(1.133)	
POST × Fit			7.216
			(9.394)
POST	−0.987	−0.632	−0.186
	(1.630)	(1.040)	(0.819)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes
Observations	8,614	8,614	8,614
Note:			*p<0.1; **p<0.05; ***p<0.01
	Standard errors clustered at the product level.		

Table 13: The impact of Q&A on ratings controlling for price, discounts, and product description lengths.

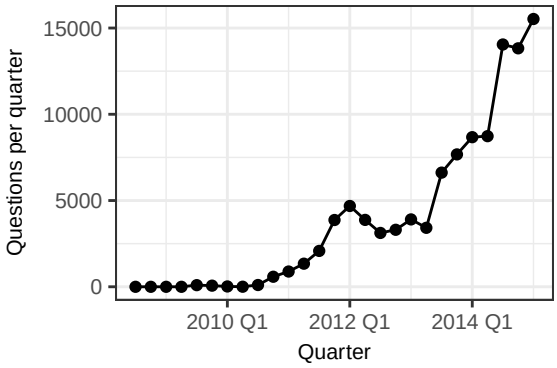
	IV (2^{nd} stage)	IV (2^{nd} stage)	IV (2^{nd} stage)
POST × Low Rating	0.120***		
	(0.041)		
POST × High Variance		0.084***	
		(0.030)	
POST × Fit			1.111***
			(0.352)
POST	−0.014	−0.028*	−0.015
	(0.012)	(0.015)	(0.013)
Review Rank	0.0001*	0.0001*	0.0001**
	(0.0001)	(0.0001)	(0.0001)
On Sale	−0.025**	−0.025**	−0.025**
	(0.011)	(0.011)	(0.011)
log(Price)	−0.141***	−0.140***	−0.142***
	(0.027)	(0.027)	(0.027)
log(Desc. Length)	0.026	0.026	0.025
	(0.016)	(0.017)	(0.016)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Observations	184,705	184,705	184,705
Note:			*p<0.1; **p<0.05; ***p<0.01
			Standard errors clustered at the product level.

Table 14: Cosine similarity around the first answer.

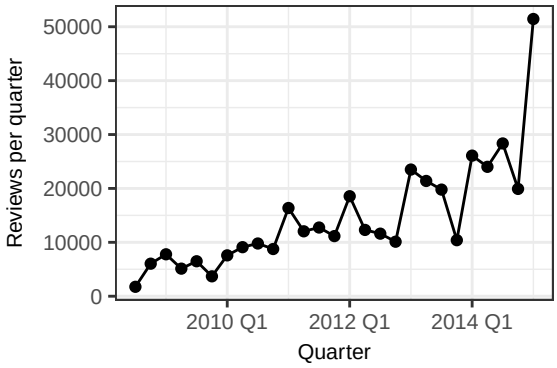
	OLS (Rating proxy)	IV (2^{nd} stage)	OLS (Var. proxy)	IV (2^{nd} stage)	OLS (Fit proxy)	IV (2^{nd} stage)
POST × Low Rating	0.003	0.018*				
	(0.005)	(0.009)				
POST × High Variance			0.008	0.013		
			(0.004)	(0.007)		
POST × Fit					−0.002	0.083
					(0.033)	(0.067)
POST	−0.003	−0.006	−0.006	−0.007	−0.003	−0.004
	(0.003)	(0.004)	(0.003)	(0.004)	(0.002)	(0.004)
Product FE	Yes	Yes	Yes			
Year-month FE	Yes	Yes	Yes			
Observations	20,432	14,741	20,432	14,741	20,432	14,741

Note: *p<0.05; **p<0.01; ***p<0.001
Standard errors clustered at the product level.

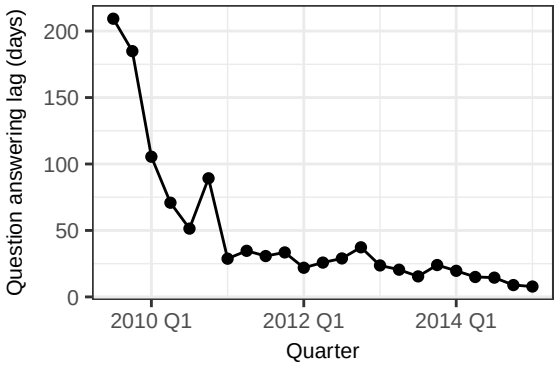
Figures



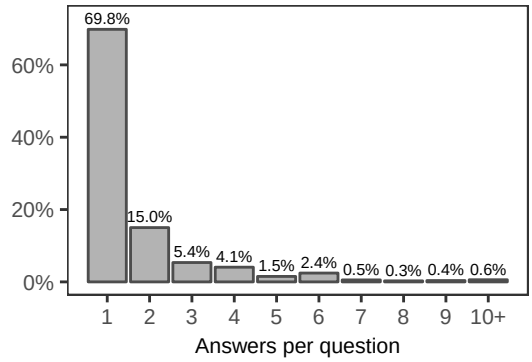
(a) Questions posted per quarter.



(b) Reviews posted per quarter.



(c) Average lag between a question and its first answer by quarter.



(d) Distribution of the number of answers per question.

Figure 1: Descriptive features of UGC: 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.

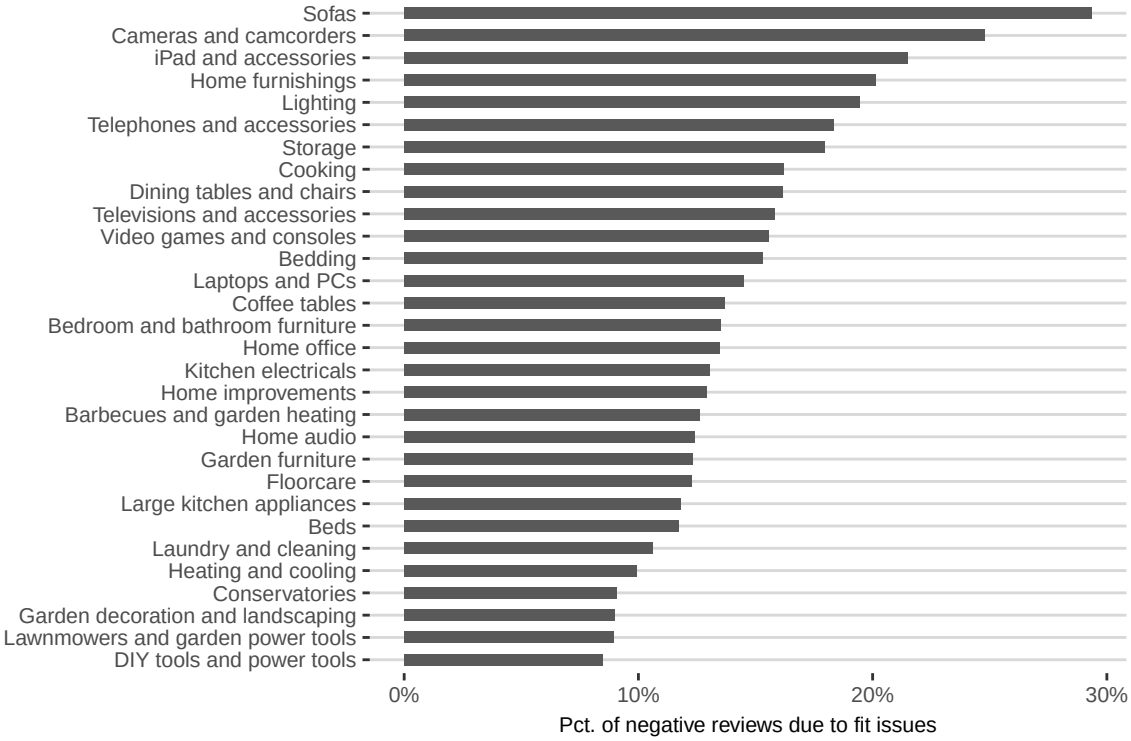


Figure 2: Percentage of negative reviews ($\leq$ 3 stars) that are due to fit-related issues by product category. (Limited to categories with at least 100 products and at least 100 reviews.)

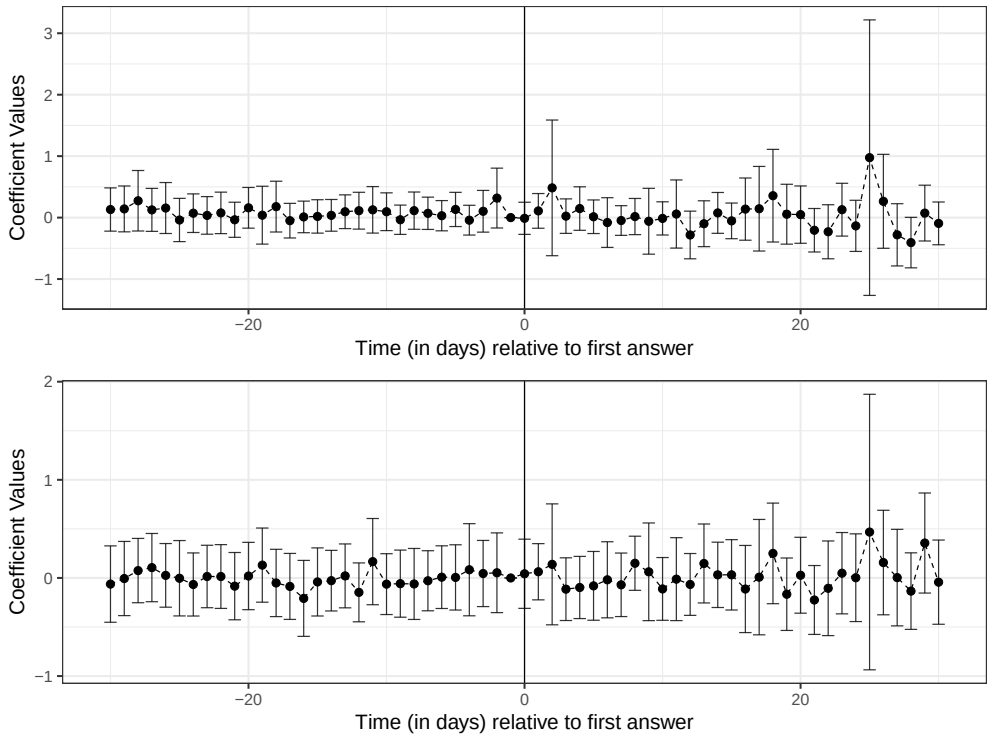


Figure 3: The evolution of review volume for products with low ratings (top) and high variance (bottom), pre and post treatment (indicated by 0), measured 30 days around the first answer. The points plot the β_k coefficient estimates from Equation 12, and the bars indicate the 95% confidence interval. We see that there are no significant irregularities around the first answer time, thus mitigating the threat of omitted variables such as promotions.

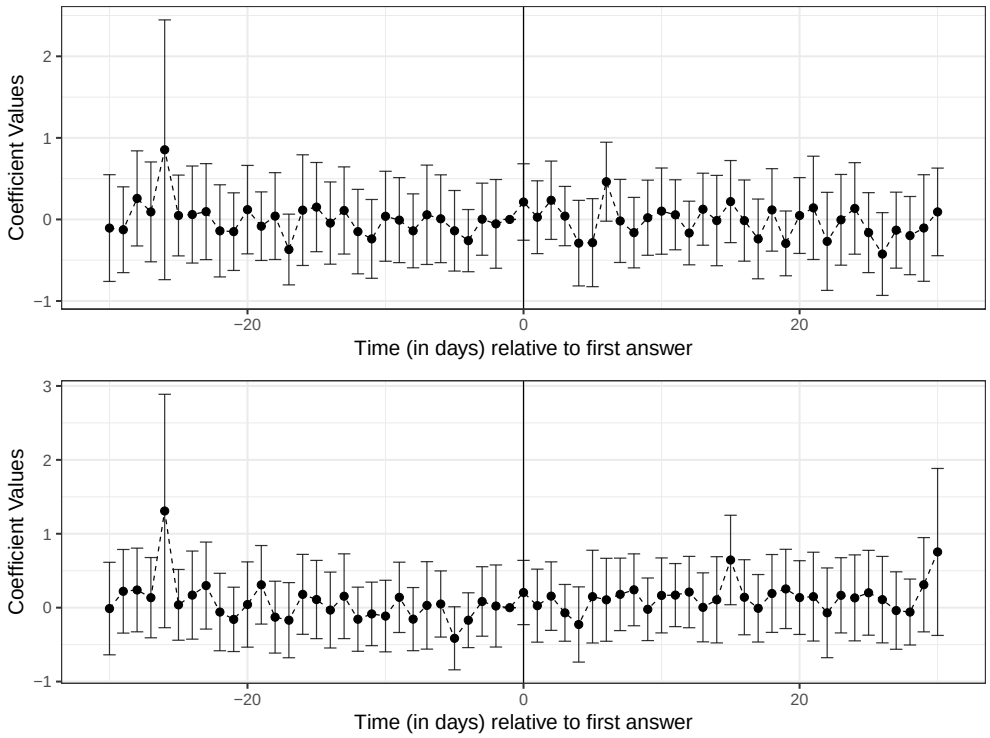


Figure 4: The evolution of pageviews for products with low ratings (top) and high variance (bottom), pre and post treatment (indicated by 0), measured 30 days around the first answer. The points plot the β_k coefficient estimates from Equation 12, and the bars indicate the 95% confidence interval. We see that there are no significant irregularities around the first answer time, thus mitigating the threat of omitted variables such as promotions.

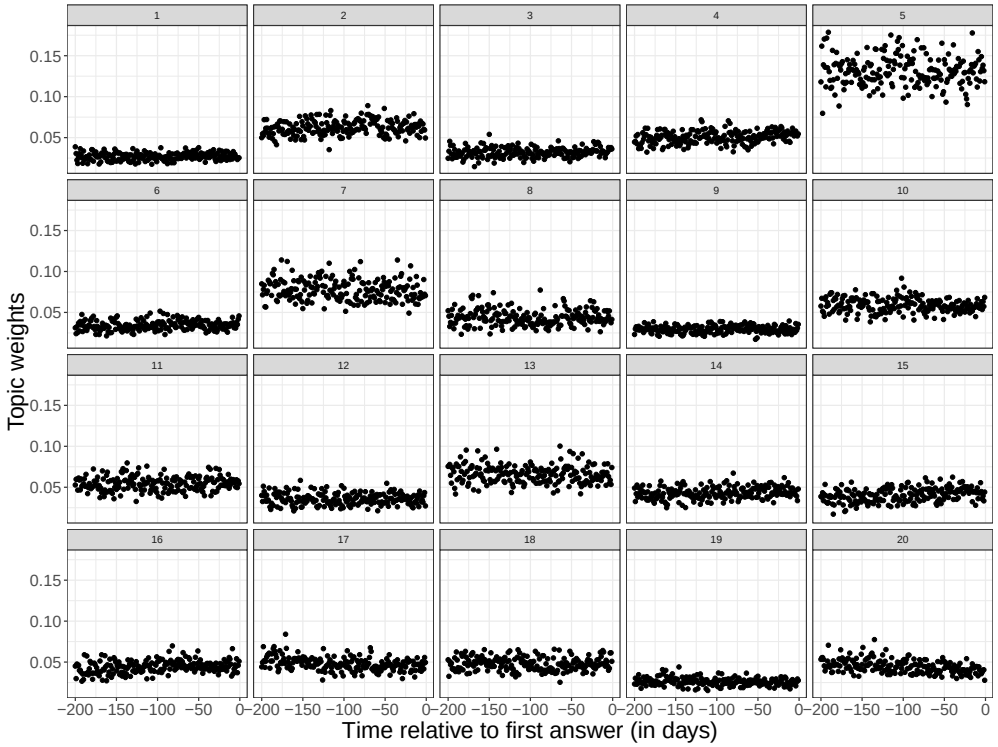


Figure 5: Plot of LDA topic weights for 20 topics, computed based on reviews that came from 0 up to 200 days prior to Q&A. We see no evidence of a change in review composition prior to Q&A arrival.

A Appendix

A.1 Tables and Figures

Table 15: LDA topics extracted from the Q&A corpus.

Topic	Highest Probability Words
Dimensions	height width depth dimensions length size
Guarantee/warranty	buy bought guarantee product warranty year
Compatibility (computers)	ipod compatible laptop windows work download
Dimensions (furniture) 1	fit sofa flat size item dimensions
Attributes (kitchen appliance)	oven grill light switch time microwave
Dimensions (furniture) 2	door side drawers left doors open
Attributes (furniture)	weight table chair chairs back seat
Installation queries (kitchen)	cooker gas included installation electric include
Compatibility (phone)	phone card sim work memory phones
Compatibility (home appliance)	box work connect record freeview internet
Attributes (lamps)	glass light pole lid lamp plastic
Attributes (power socket)	cable usb plug battery socket power
Attributes (entertainment system)	work player dvd play remote samsung
Dimensions (bed)	bed mattress size fit base double
Queries (home appliance)	fridge freezer free dryer long wash
Dimension (furniture) 3	wall unit shelves shelf top fit
Instructions (camera/printer)	camera printer print clock ink set
Description discrepancies	product description confirm question correct states
Color/finish (furniture)	colour made white black wood match
Attributes (home appliance)	machine water washing make hot filter

Table 16: LDA topics extracted from the review corpus.

Topic	Highest Probability Words
Quality (vacuum)	easy great good cleaner clean product
Quality (electronics)	sound good great clock quality set
Quality (phone/camera)	phone easy set good camera features
Quality (home appliance)	kettle water good iron machine toaster
Quality (bedclothes)	bed duvet comfortable warm mattress pillows
Value for money 1	good job easy product money price
Product instructions	put instructions wall screws holes fit
Product returns	store item product service delivery back
Replacement	bought years printer replace buy good
Gifts	bought great room son loves year
Quality (clothes line)	put clothes easy good cover sturdy
Dimensions (storage unit)	storage easy small space put unit
Value for money 2	product money great good recommend excellent
Value for money (negative)	it's bit reviews price cheap buy
Quality (garden tools)	good job cut light easy small mower
Discounts	price good quality great bargain sale
Quality (lighting)	light colour nice lovely lamp room
Assembly instructions	easy good put table assemble money
Quality (kitchen appliance)	easy clean great cooker microwave food
Quality (poor)	quality bin back poor plastic cheap

Table 17: Review volume following the first answer.

	OLS (Rating proxy)	OLS (Var. proxy)	OLS (Fit proxy)
POST × Low Rating	−0.037		
	(0.033)		
POST × High Variance		0.020	
		(0.037)	
POST × Fit			−0.157
			(0.105)
POST	0.089*	0.074	0.085*
	(0.050)	(0.054)	(0.047)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes
Observations	37,863	37,863	37,863
Note:			*p<0.1; **p<0.05; ***p<0.01
			Standard errors clustered at the product level.

Table 18: Pageview volume following the first answer.

	OLS (Rating proxy)	OLS (Var. proxy)	OLS (Fit proxy)
POST × Low Rating	0.318		
	(0.468)		
POST × High Variance		0.607	
		(0.479)	
POST × Fit			0.710
			(0.930)
POST	−0.424	−0.506	−0.349
	(0.410)	(0.408)	(0.349)
Product FE	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes
Weekday FE	Yes	Yes	Yes
Observations	20,209	20,209	20,209

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

A.2 Mathematical Appendix

In this appendix we present a stylized model that captures the essence of how the presence of informative Q&A affects consumer decision making and product ratings in settings with consumer fit uncertainty.

Consumer Side. We begin by modeling the consumer side. Understanding how the presence of Q&A affects consumer behavior is essential in order to understand the impact of Q&A on average product ratings.

A focal consumer contemplates whether to purchase a product. The consumer possesses perfect knowledge about every attribute of the product, except one. The unknown attribute can take one of two values, a "good" value resulting in positive product utility g and a "bad" value resulting in negative utility $-b$ ($g, b \geq 0$). Both utilities g, b include the disutility of price. All utilities, as well as the definition of what is "good" and "bad", might differ from one consumer to the next. Therefore, our model is general enough to encompass both quality and fit-related attributes. In the latter case, "good" and "bad" have subjective interpretations. For example, "good (bad)" might mean "the dimensions of this product fit (do not fit) through my apartment's door" or "the lens is (is not) compatible with my camera."

Let us denote by α the prior probability that the unknown attribute will take the "bad" value; α is thus the probability of "bad fit" or fit mismatch. In the absence of any additional information, the consumer's expected utility is $g-\alpha(g+b)$. The consumer purchases if and only if $g-\alpha(g+b) > 0$ or, equivalently, if $\alpha < g/(g+b)$. If the consumer purchases, with probability α she experiences negative post-purchase utility, i.e. *regrets* the purchase. We assume that if the consumer experiences positive (negative) post-purchase utility she posts a positive (negative) product review. Assuming that a positive review is equivalent to a rating of "1" and a negative review equivalent to a rating of "0", the average product rating is equal to one minus the average probability of post-purchase regret among its purchasers.[21]

Let us now assume that the consumer asks a question about the value of the unknown attribute (by posting a question at a Q&A forum) and receives back an answer. The answer can be positive, meaning "the attribute is good" or negative, meaning, "the attribute is bad". We assume that the answer is correct with probability $p \geq \frac{1}{2}$. Thus, p denotes the quality of information. Denote by π_+, π_- the posterior probabilities that the unknown attribute has the "bad" value given positive (+) or negative (-) answers respectively. According to standard Bayesian inference, it is:

[21] The model can be extended to a multi-valued rating scale $1, 2, ..., n$ by defining a correspondence between post-purchase utilities $u_1, u_2, ..., u_{n-1}$, where $u_i < u_{i+1}$, such that consumers post rating i if they experience post-purchase utility $u_{i-1} < u \leq u_i$ plus the obvious corner cases. The precise thresholds u_i may differ among consumers. Such a mapping retains the key properties that drive our stylized model, i.e. average ratings are positively related to average post-purchase utility and negatively related to the probability of fit mismatch among purchasers.

$$\pi_+ = \frac{Pr[+|b]Pr[b]}{Pr[+]} = \frac{(1-p)\alpha}{(1-p)\alpha + p(1-\alpha)} \qquad \pi_- = \frac{Pr[-|b]Pr[b]}{Pr[-]} = \frac{p\alpha}{p\alpha + (1-p)(1-\alpha)}$$

It is easy to show that:

- π_+ is monotonically decreasing with p and ranges from α (for $p = \frac{1}{2}$) to 0 (for $p = 1$)
- π_- is monotonically increasing with p and ranges from α (for $p = \frac{1}{2}$) to 1 (for $p = 1$)
- $\pi_+ \leq \alpha \leq \pi_-$ for all $p \geq \frac{1}{2}$

The consumer's expected utility from purchase, given answer $s \in \{+, -\}$, is equal to:

$$u_s = (1 - \pi_s)g + (\pi_s)(-b) = g - \pi_s(g + b)$$

The consumer purchases if and only if $u_s > 0$. There are two cases:

Case I: (Optimistic consumers) $g - \alpha(g + b) > 0$, or equivalently $\alpha < g/(g + b)$. In this case the consumer would always purchase the product on the basis of her prior beliefs. If we add the option of asking questions, the consumer purchases the product if either: 1) she receives a positive answer of any informativeness, or 2) she receives a negative answer of low informativeness, such that her posterior beliefs remain close to the prior. However, she does not purchase the product if she receives a negative answer whose informativeness p is sufficiently high, such that $\pi_- \geq g/(g + b)$. The consumer's probability of regret conditional on purchase, is equal to α if there is no Q&A or if there is Q&A, as long as the answer's informativeness remains relatively low. The probability of regret, conditional on purchase, *decreases* to π_+ (recall that $\pi_+ \leq \alpha$) as soon as the answer's informativeness p crosses the threshold above which the consumer buys only if she receives a positive answer; π_+ is a declining function of p and converges to zero as p tends towards 1, i.e. as answers to questions become perfectly reliable. In that limiting case, consumers who choose to purchase in the presence of Q&A never experience fit mismatch and post only positive ratings.

Case II: (Pessimistic consumers) $g - \alpha(g + b) \leq 0$, or equivalently $\alpha \geq g/(g + b)$. In this case, the consumer would not purchase the product on the basis of her prior beliefs. If we add the option of asking questions, the consumer only purchases the product if she receives a positive answer whose informativeness p is sufficiently high, such that $\pi_+ \leq g/(g+b)$. If the consumer purchases, the probability of post-purchase regret is π_+; as above, the probability of regret goes to zero in the limiting case of perfectly reliable answers.

The conditions that determine whether Case I or Case II applies depend on both α and the ratio $g/(g + b)$. Case I applies when either α is small or $g/(g + b)$ is large. Case II applies when α is large or $g/(g + b)$ is small. The ratio $g/(g + b)$ captures the relationship between the utility of a

match g and the disutility of a mismatch b. In settings where the consequences of a mismatch are not very severe (e.g. when product prices are low and/or products are easy to return), b is likely to be small and $g/(g+b)$ large. Conversely, in settings where the consequences of a mismatch are more severe (e.g. high prices, difficult to return products, bad fit causes damage to property or health) b is likely to be large and $g/(g+b)$ small.

In summary:

1. The presence of Q&A affects consumer decision making only if answers are sufficiently informative (i.e. if p is sufficiently high).

2. The ability to ask a question and receive a sufficiently informative answer about an unknown fit-related product feature has the following effect on consumer decision making:

   (a) In settings where the prior probability of bad fit is low or the consequences of fit mismatch not severe (Case I), it discourages consumers who would otherwise be making a mistake from purchasing the product if the answer indicates that the product may not be a good fit for them.

   (b) In settings where the prior probability of bad fit is high or the consequences of fit mismatch severe (Case II), it encourages consumers who would otherwise be reluctant to purchase the product if the answer indicates that the product may be a good fit for them.

Observe that, in our model, average consumer utility and average ratings are linear transformations of one another. Specifically, average consumer utility $u = g - \pi(g+b)$ corresponds to average product rating $r = 1 - \pi$, which gives:

$$r = \frac{1}{g+b}u + \frac{b}{g+b}$$

Product Side. We now turn our attention to the product side. For any given product, we assume that there are multiple prospective consumers. We, further, assume that a fraction $1-\epsilon$ of consumers (we will call them the "informed" consumers) have perfect information about the product and purchase it, knowing that it serves their needs. These consumers always post positive reviews. The remaining consumers behave like the focal consumer we analyzed above. We will call those consumers the "uninformed" consumers.

Each uninformed consumer may care about different unknown attributes and may have different notions of what constitutes "good" and "bad" states of those attributes. On aggregate, we assume that the focal product is a bad *ex-post* fit for a fraction ω of uninformed consumers and a good fit for the rest. However, uninformed consumers do not have precise fit information *ex-ante* and, as discussed above, make decisions assuming a prior probability of bad fit equal to α.

Note that there is no inconsistency in assuming different values for α and ω. Whereas the value of α reflects the distribution of product attributes on the market, ω reflects the distribution

of consumer tastes for those attributes. For example, consider the case of portable hard drives that can be compatible with PC only or Mac only. Assume that 70% of portable hard drives on the market are compatible with PC only and 30% compatible with Mac only. If the drive's compatibility is the unknown attribute, a PC user would be justified in assuming $\alpha = 0.3$ whereas a Mac user would be justified in assuming $\alpha = 0.7$. Assume, now, that 90% of consumers are PC users and 10% are Mac users. A PC-compatible hard drive is, thus, a bad fit for a fraction $\omega = 0.1$ of consumers, whereas a Mac-compatible hard drive is a bad fit for a fraction $\omega = 0.9$.

Under the above assumptions, and assuming that Q&A is informative enough (i.e. that p is sufficiently high) to affect consumer behavior:

1. If $\alpha < g/(g+b)$, such that Case I applies:

   (a) Without Q&A, all uninformed consumers purchase. A fraction ω will experience bad fit and will post negative reviews. The average ratings of uninformed consumers will then be $1-\omega$ and the average ratings of all consumers (informed plus uninformed) $1-\epsilon+\epsilon(1-\omega)$.

   (b) With Q&A, with probability p a fraction ω of uninformed consumers (the fraction for whom the product is a bad fit) will receive a (correct) negative answer and will not purchase. With probability $1-p$ that same fraction will receive a (wrong) positive answer and will purchase, resulting in bad fit and negative reviews. With probability p the remaining fraction (the fraction for whom the product is a good fit) will receive a (correct) positive answer and will purchase, resulting in positive reviews. With probability $1-p$ that same fraction will receive a (wrong) negative answer and will not purchase. The average ratings of uninformed consumers will thus be $p(1-\omega)/[(1-p)\omega+p(1-\omega)]$ and the average ratings of all consumers $(1-\epsilon+\epsilon p(1-\omega))/[1-\epsilon+\epsilon((1-p)\omega+p(1-\omega))]$.

2. If $\alpha \geq g/(g+b)$, such that Case II applies.

   (a) Without Q&A, no uninformed consumers purchase. Informed consumers always post positive ratings, therefore, the average product ratings will be 1.

   (b) With Q&A, the effect will be identical to case 1(b) above.

Without loss of generality, we assume that a fraction γ of uninformed consumers have prior beliefs α such that Case I applies and the rest have prior beliefs such that Case II applies. Then, combining Cases 1 and 2 above, we conclude that:

- Without Q&A (Cases 1(a) and 2(a)), the average rating of the product will be $(1-\epsilon+\epsilon\gamma(1-\omega))/(1-\epsilon+\epsilon\gamma)$. Using elementary comparative statics we can show that this is a decreasing function of ϵ, ω and γ, that is, ratings are lower when:

  1. there are many uninformed consumers (high ϵ), that is, the product exhibits a higher fit uncertainty, and

2. the product is a bad fit for a large fraction of consumers (high ω), that is, the product caters to niche tastes that do not coincide with the mainstream [22], and

3. many uninformed consumers are optimistic about the probability of a good fit and choose to purchase in the presence of fit uncertainty (high γ); as previously discussed, this happens when most products of this category are a good fit for most consumers (such that most consumers have a low α) and/or when then impact of bad fit is not very severe relative to the utility of a good fit.

- An interesting nuance of this result is that *high* average ratings may indicate one or more of the following conditions:

  1. there exist few uninformed consumers (low ϵ)

  2. the product is a good fit for a lot of consumers (low ω)

  3. most uninformed consumers are pessimistic and choose to not purchase; as previously discussed, this happens when the probability of fit mismatch is high for this product category (high α) and/or the impact of bad fit is severe relative to the utility of a good fit

- With sufficiently informative Q&A (Cases 1(b) and 2(b)), γ does not matter and the average rating of the product will be $(1-\epsilon+\epsilon p(1-\omega))/[1-\epsilon+\epsilon((1-p)\omega+p(1-\omega))]$. Using elementary comparative statics we can show that this is an increasing function of p and a decreasing function of ϵ and ω. Average ratings are higher, the higher the informativeness of Q&A and the lower the fraction of 1) uninformed consumers and 2) consumers for which the product is not a good fit.

The crispest intuitions are obtained when Q&A answers are always correct ($p=1$). Average ratings with Q&A are then equal to 1 for all ω, since consumers become perfectly informed and purchase if and only if the product is a good fit for them. The ratings increase due to Q&A is then simply $1-(\textit{Ratings without Q\&A})= \epsilon\gamma\omega/(1-\epsilon+\epsilon\gamma)$.[23] The latter is an increasing function of ϵ, ω and γ. The ratings increase is also inversely proportional to the ratings without Q&A, i.e. the lower these ratings, the higher the increase.

To reiterate, the positive impact of Q&A on average ratings of individual products is highest for products that

1. exhibit fit uncertainty for many consumers, and

2. are not a good fit for many consumers, and

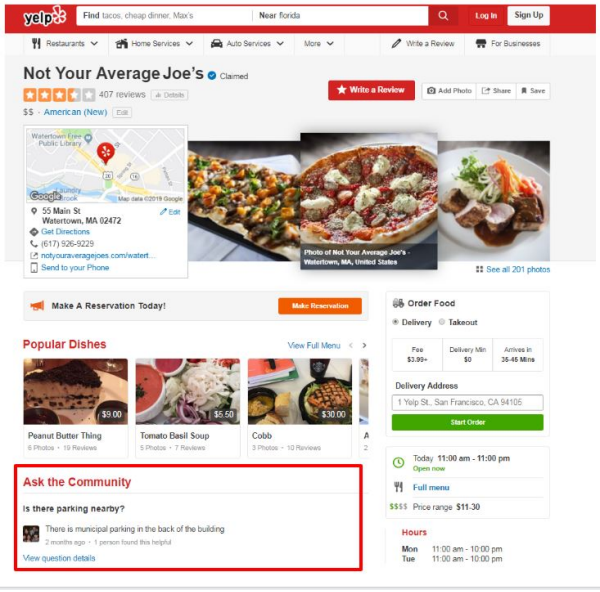
[22] Assuming that most computer users are PC users, an external hard drive that is only compatible with Apple computers would be an example of such a niche product.

[23] This expression roughly corresponds to m_j in our empirical model.

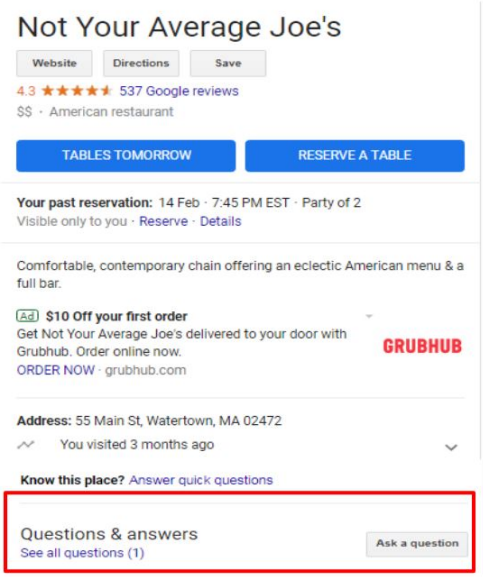
3. belong to product categories where many uninformed consumers are optimistic about the probability of a good fit and choose to purchase in the presence of fit uncertainty; as previously discussed, this happens when most products of this category are a good fit for most consumers and/or when the impact of bad fit is not very severe relative to the utility of a good fit.

When $p < 1$ the ratings increase due to Q&A is generally lower and may even become negative in settings where γ is close to 0. Such settings are characterized by pessimistic uninformed consumers who have very unfavorable priors about product fit or are faced with severe consequences of fit mismatch (i.e. fall in Case II of the consumer model). In the absence of Q&A, pessimistic uninformed consumers do not purchase; only informed consumers purchase and post positive ratings. The presence of Q&A may convince uninformed consumers to purchase if they receive a positive answer. However, because this answer may be wrong with some probability, some uninformed purchasers will experience bad fit and will post negative ratings, thus lowering the (previously perfect) average. In the paper we assume that p is sufficiently close to one (and/or γ sufficiently high) for such effects to *not* occur. Our empirical analyses are consistent with these assumptions; we find no statistically significant evidence of average ratings *declining* after questions are answered.

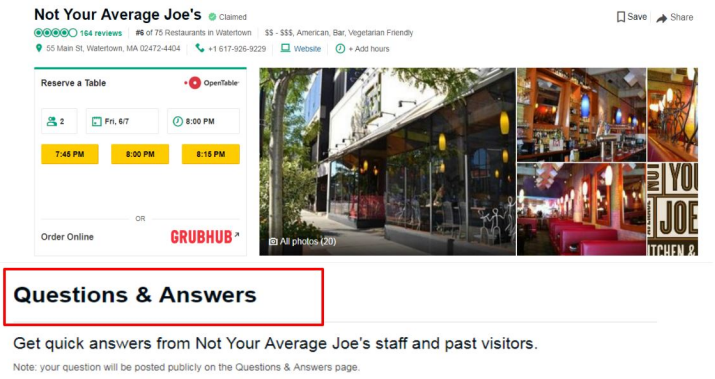
B Additional Tables and Figures (Web Appendix)



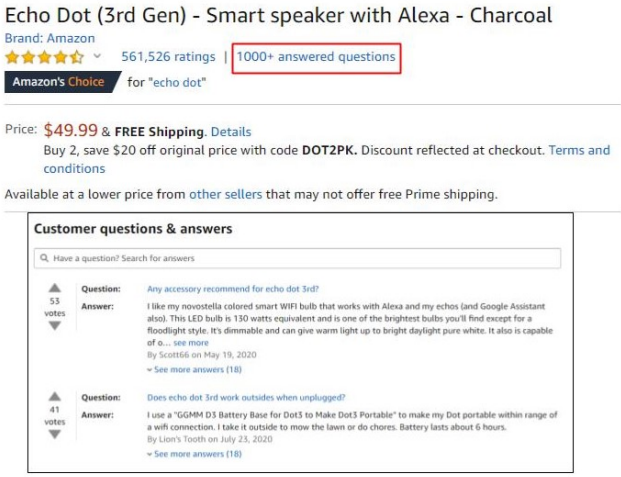
(a) Yelp.



(b) Google reviews.

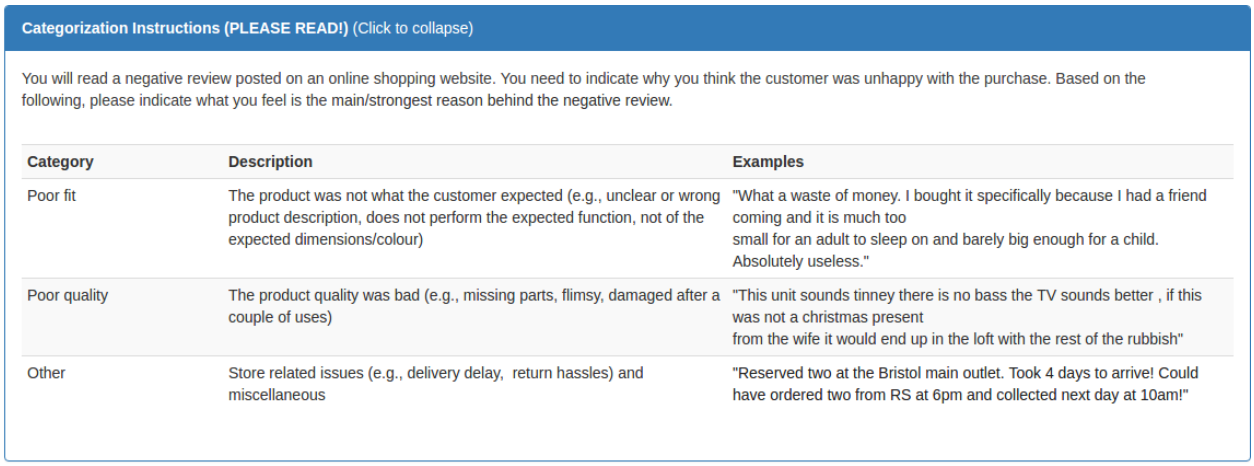


(c) TripAdvisor.



(d) Amazon.

Figure 6: Q&A technology on different platforms.



Categorization Instructions (PLEASE READ!) (Click to collapse)

You will read a negative review posted on an online shopping website. You need to indicate why you think the customer was unhappy with the purchase. Based on the following, please indicate what you feel is the main/strongest reason behind the negative review.

Category	Description	Examples
Poor fit	The product was not what the customer expected (e.g., unclear or wrong product description, does not perform the expected function, not of the expected dimensions/colour)	"What a waste of money. I bought it specifically because I had a friend coming and it is much too small for an adult to sleep on and barely big enough for a child. Absolutely useless."
Poor quality	The product quality was bad (e.g., missing parts, flimsy, damaged after a couple of uses)	"This unit sounds tinney there is no bass the TV sounds better , if this was not a christmas present from the wife it would end up in the loft with the rest of the rubbish"
Other	Store related issues (e.g., delivery delay, return hassles) and miscellaneous	"Reserved two at the Bristol main outlet. Took 4 days to arrive! Could have ordered two from RS at 6pm and collected next day at 10am!"

Figure 7: Screenshot of survey shown to workers on MTurk.

Table 19: Examples of reviews indicating fit and quality issues.

Fit issues	Quality issues
1. What a waste of money. I bought it specifically because I had a friend coming and it is much too small for an adult to sleep on and barely big enough for a child. Absolutely useless. (3 stars)	1. I suppose it's true that you get what you pay for. It's light and compact, no problem to set up, but the sound quality is very poor. I call it Tin Lizzie. (2 stars)
2. Bought for an occasional put-u-up for the grandchild on sleep over - suitable for small child, folding away to make a convenient bed chair. NOT SUITABLE FOR 10+. As it's close to the ground and contains no metal or sharp pieces, the child cannot hurt itself if rolls out of bed!! For what it is could be a lower price. (3 stars)	2. Disappointed to say the least. The beads were not aligned and the bar at the top is just a strip of cheap wood that is bent. (1 star)
3. The lead works great on some tomtoms with the larger USB connector but the adapter supplied doesn't work on the smaller USB connector. Have tried another connector, still no luck, so lead stuck in drawer now!!!! (1 star)	3. Both the store manager and I tried to fit the case, on the date of purchase (it was the appropriate design for my iPod), but the two halves did not fit together. The case was flimsy too, so I don't know how much protection it would have afforded my device anyway. I received my money back there and then. (1 star)
4. this item is CREAM & black (NOT white & black)we ordered this online & it was indeed very comfortable & well made. However, it was cream & black in colour, so we returned it and had a look at another in the store that was exactly the same. The manager said he would feed the colour problem back to head office. We will look elsewhere for a black & white beanbag !so - in summary - if you are after a CREAM and black football bean bag then you would be very happy with this product. (2 stars)	4. My son used these twice then they stopped working properly. (1 star)

Table 20: Flexible definition of holdout: The impact of Q&A on ratings using low pre-treatment ratings (≤ 4) as a proxy for fit uncertainty.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1^{st} stage)	IV	IV (bins)
POST × Low Rating	0.243^{***}	0.183^{***}		0.204^{***}	
	(0.022)	(0.027)		(0.036)	
POST × Hold-out Rating			-0.918^{***}		
			(0.025)		
POST	-0.045^{***}	-0.022^{**}	4.121^{***}	-0.025^{**}	
	(0.009)	(0.011)	(0.110)	(0.011)	
POST × Rating $\in [2, 3]$					0.557^{***}
					(0.107)
POST × Rating $\in (3, 4]$					0.128^{***}
					(0.031)
POST × Rating $\in (4, 5]$					-0.018^{*}
					(0.011)
Review Rank	0.0002^{***}	0.0001^{*}	-0.0001^{*}	0.0001^{*}	0.0001^{*}
	(0.0001)	(0.0001)	(0.00005)	(0.0001)	(0.0001)
Product FE	Yes	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes	Yes
F Statistic			455.98		
Observations	345,168	225,182	225,182	225,182	225,182

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 21: Flexible definition of holdout: The impact of Q&A on ratings using high pre-treatment rating variance (≥ 1) as a proxy for fit uncertainty.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1^{st} stage)	IV
POST × High Variance	0.157***	0.124***		0.153***
	(0.015)	(0.017)		(0.025)
POST × Hold-out Variance			0.611***	
			(0.012)	
POST	−0.060***	−0.042***	−0.155***	−0.053***
	(0.010)	(0.011)	(0.015)	(0.013)
Review Rank	0.0001***	0.0001*	0.00001	0.0001
	(0.00004)	(0.0001)	(0.0001)	(0.0001)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			823.99	
Observations	345,168	225,182	225,182	225,182

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 22: Flexible definition of holdout: The impact of Q&A on ratings using the pre-treatment fraction of review mentioning fit issues as a proxy for fit uncertainty.

	OLS (Full sample)	OLS (Excl. hold-out)	IV (1^{st} stage)	IV
POST × Fit	1.933***	1.940***		1.722***
	(0.150)	(0.264)		(0.348)
POST × Hold-out Fit			1.440***	
			(0.041)	
POST	−0.039***	−0.029**	−0.007***	−0.025**
	(0.010)	(0.012)	(0.001)	(0.012)
Review Rank	0.0002***	0.0002**	−0.00000*	0.0001**
	(0.00005)	(0.0001)	(0.00000)	(0.0001)
Product FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
F Statistic			418.86	
Observations	345,168	225,182	225,182	225,182

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at the product level.

Table 23: LDA topics extracted from the pre-Q&A reviews.

Topic	Highest Probability Words
Furniture 1	table seat present glad black
Value for money 1	excellent money value good product
Quality 1	set feature delight read pretty
Phone	recommend sound phone work highly
Value for money 2	good great easy look price
Replacement	old year bought replace purchase
Storage	small easy fit space storage
Furniture 2	bed comfort chair bought mattress
Accessories	design star rang push piece
Returns	return better connect review work
Usage	easy use said work simple
Holidays	iron lid bag bin christmas
Quality 2	light look love colour nice
Furniture 2	price love table great bought
Kitchen appliance	use heat water cook kettle
Household appliance	use clean floor cleaner vacuum
Instructions	instruct wall screw bit drill
Outdoor equipment	machine product price garden
Clocks	clock keep cheap time real
Furniture 3	cover door easy plenty heavy