The Role of Attributes in Product Quality Comparisons

Felipe Moraes*
f.moraes@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Jie Yang
jiy@amazon.com
Amazon
Seattle, USA

Rongting Zhang
rongtz@amazon.com
Amazon
Seattle, USA

Vanessa Murdock
vmurdock@amazon.com
Amazon
Seattle, USA

ABSTRACT

In online shopping quality is a key consideration when purchasing an item. Since customers cannot physically touch or try out an item before buying it, they must assess its quality from information gathered online. In a typical eCommerce setting, the customer is presented with seller-generated content from the product catalog, such as an image of the product, a textual description, and lists or comparisons of attributes. In addition to catalog attributes, customers often have access to customer-generated content such as reviews and product questions and answers.

In a crowdsourced study, we asked crowd workers to compare product pairs from kitchen, electronics, home, beauty and office categories. In a side-by-side comparison, we asked them to choose the product that is higher quality, and further to identify the attributes that contributed to their judgment, where the attributes were both seller-generated and customer-generated. We find that customers tend to perceive more expensive items as higher quality but that their purchase decisions are uncorrelated with quality, suggesting that customers seek a trade-off between price and quality when making purchase decisions. Crowd workers placed a higher value on attributes derived from customer-generated content such as reviews than on catalog attributes. Among the catalog attributes, brand, item material and pack size[1] were most often selected. Finally, attributes with a low correlation with perceived quality are nonetheless useful in predicting purchases in a machine-learned system.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Users and interactive retrieval**; **Crowdsourcing**.

KEYWORDS

Product quality; attribute comparison; online reviews

*Work conducted while the first author was an intern at Amazon.
[1]the number of items in a multipack, e.g., a two-cable pack.

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377956

ACM Reference Format:
Felipe Moraes, Jie Yang, Rongting Zhang, and Vanessa Murdock. 2020. The Role of Attributes in Product Quality Comparisons. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3343413.3377956

1 INTRODUCTION

In online shopping, customers seek items that are not only relevant, but also high quality, useful, appealing, and good value. For example, if the customer asks to buy noise-cancelling headphones, many products could be considered relevant, but the price range is large ($50 to $350 or more) and there is a substantial difference in the effectiveness of the noise cancellation, the sound quality, the comfort of the ear covers, the weight of the headphones, etc. A purchase decision is typically made by weighing all these factors.

Like the concept of *relevance*, people do not agree on what it means for a result to be *high quality*, *useful*, or *appealing*. These qualifiers are contextual, if not personal, and difficult to pin down. It is the job of a recommender system to use the many sources of information from the seller or manufacturer of the product, from experts who recommend the products on sites like Wirecutter,[2] and from customers who have purchased and used the products to make the best recommendation.

In this paper we examine how users understand an item to be *higher quality*. In a crowdsourced study, we asked 420 crowd workers to compare 946 product pairs from kitchen, electronics, home, beauty and office categories. We define *comparable products* to be products that satisfy a similar need, even if they differ in their attributes. We sampled search sessions from the logs of a large eCommerce engine, and chose product pairs where one product was purchased, and the other was clicked in the same session. We presented the products side-by-side and asked the crowd workers to determine which product was higher quality.

Customers have multiple sources of information available when doing product comparisons online. The sellers of the products provide basic information about the product such as an image, description, title, and lists of attributes such as its size, weight, and other specifications. Many products are reviewed by other customers who have already purchased and used the product, and some sites offer customer question-answer pages, where customers who are considering a purchase can ask a question to be answered by another customer. It is not clear which sources of information are

[2]https://thewirecutter.com/ visited October 2019.

most valuable to customers making purchase decisions. To understand this, the side-by-side interface included the top five customer reviews (which may be a mix of positive and negative reviews) and the top three product Q&A. We recorded crowd worker clicks to expand the reviews and product Q&A.

Finally, to understand which attributes customers consider when making purchase decisions, we presented a list of seller-generated attributes and attributes extracted from the customer reviews and product Q&A, and asked the crowd workers to identify which attributes they considered when determining the product quality.

We find that the price of the item correlates highly with product quality, but product quality does not correlate with the actual purchase behavior from the logs. Similarly, star-ratings of the reviews that have been widely used to approximate customer preferences, do not correlate with perceived product quality. Among seller- and customer-generated content, crowd workers selected more customer-generated attributes in the quality assessment, yet most of the workers did not choose to expand the reviews and read them for the assessment. Among the seller-generated attributes, crowd workers identified brand name, item material and pack size most often. Specific to customer-generated content, a wider range of attributes were considered useful and differed by product categories, showing that customer-generated content provides a rich and diverse source of information for quality assessment. Finally, we also find that crowd workers who self-identified as frequent buyers have higher agreement on which attributes are important.

The remainder of this paper is organized as follows: we present related work in Section 2; we describe our data collection in Section 3, before turning to the experimental setup of the crowdsourced study in Section 4. The results are outlined and discussed in Section 5. Finally, we conclude with an overview of future directions in Section 6.

2 RELATED WORK

There has been considerable research in determining item quality, in both "real world" and online settings. In the eCommerce domain, the aim is often to determine which products are valid substitutes of another product, and whether customer behaviors such as purchases and ratings can be used as a proxy for quality, similarity or substitutability. Our work conducts a crowdsourced study to ask people about quality, and the utility of customer-generated content explicitly, and then determines whether it correlates to customer behaviors in the logs of a large eCommerce engine.

2.1 Item Quality Offline and Online

Item quality is a multifaceted property, for which it is difficult to establish a holistic definition. Relevant dimensions are, for example, flawlessness, durability, appearance, and distinctiveness [27]. It is generally agreed that quality is a perceptual and conditional property that consists of item features that meet customers' expectations affected by situational factors [12, 18]. Quality is, therefore, mostly studied in the notion of "perceived quality" and quantified by questionnaires [11, 20, 21].

Product quality has been extensively studied in offline shopping settings, but not in online settings. In offline settings, it has been found that both intrinsic attributes, i.e., the integral components of the physical product (e.g., material), and extrinsic attributes that are not part of the physical product (e.g., brand name), are indicators of perceived quality. In particular, correlation has been found between price and perceived quality, as well as between brand name and perceived quality for products [23]. Apart from these attributes, other product attributes that affect quality perception can differ significantly across product types [21]. In online shopping environments, since customers cannot assess product quality physically, it is unknown whether attributes that are relevant offline remain relevant in online settings. For the same reason, customers have the tendency to rely on customer-generated content for quality assessment, yet it remains a question how much this type of information plays a role in quality assessment.

While little work has studied product quality online, some work has studied the quality of user experiences (QoE) in image and video applications [26]. QoE is also known as user-perceived quality that ties together user perceptions and expectations with respect to the application; it is, therefore, close to the definition of quality for products. QoE is usually measured by subjective assessment that involve real users in an in-house or crowdsourced setting [8, 10]. Results from quality assessment have been found to be consistent across participants in a variety of applications, including video streaming, Web surfing, file download, images, and cloud gaming [9]. This allows for the construction of golden labels that can be used to train machine learning models for automated quality modeling [7, 13, 17, 19]. Compared with this work, little is known about whether product quality perception is consistent across customers, which we investigate in our study. In addition, we contribute insights into how much customers agree with product attributes relevant for quality assessments, and how such agreement is influenced by customers' shopping experience.

2.2 Product Quality vs. Similarity and Substitutability in eCommerce

In eCommerce, *quality* has been a largely neglected concept compared with *relevance* and *preference*. Related work can be found on modeling and predicting customer ratings to products [6, 22, 25, 28] for product recommendation. Ratings have been mostly viewed as a proxy for customer preferences that involve customer interests, functional needs for products, and perceived product quality. In addition, recent work has shown that customer ratings are to a large extent influenced by algorithmic and self-selection biases [4, 24]. Ratings, therefore, might not be sufficient to support the development of product recommendation. For this reason, there has been a growing interest in finding out the reasons behind customers' rating or purchase decisions, by uncovering the relationships between products such as similarity [29] and substitutability [14, 30]. Our work extends the literature by contributing insights into the relationship between product quality and customer behaviors such as ratings and purchases.

For recommendation, understanding the relative quality between a pair of products is useful for recommending a product similar to a given one (e.g., the one being browsed) but of higher quality. Work on similarity and substitutability serves as a basis for our work, since customers usually determine product quality by comparing similar or substitutable products.

McAuley et al. [14, 15] introduce a machine learning approach to infer the substitutability relationship between products using the text from product reviews and descriptions or the images of products. Product pairs that are viewed together where one is purchased are taken as the ground truth for model training. We take a similar approach to construct comparable pairs of products. To ensure the comparability of products in terms of function, we further filter the pairs by other heuristics (e.g., same category). In addition, we adopt a user-centric study to understand which attributes are deemed important by customers.

In a recent paper, Trattner and Jannach [29] study item similarity for recommendation. Similar to this work, we rely on human judgment as the gold standard. Unlike our work, they do not ask human workers to identify the attributes that contribute to workers' judgment. We note that attributes collected by workers are important for understanding the reasons behind the judgment as well as for training explainable recommenders [31], be they similarity-based or quality-based.

3 DATA COLLECTION

The data collection process has two stages: 1) constructing product pairs that are comparable, and 2) extracting product attributes from the product catalog and customer-generated content.

3.1 Mining Comparable Products Pairs

It is often easier for customers to judge the relative quality of two items rather than the absolute quality of one item [2]. For this reason, we construct pairs of comparable products for quality assessment. We consider two products comparable if they satisfy similar customer needs, even if they differ by price, by quality or by specific attributes. We consider only products from kitchen, electronics, home, beauty, and office supplies categories.

3.1.1 Candidate Pairs Generation. We mined pairs from the customer interaction logs of a large eCommerce engine over a six-month period from April 16th to September 30th, 2018. We segmented the logs into shopping sessions, where sessions are all customer activity leading to a purchase, or to 30 minutes of inactivity. We selected candidate product pairs (i, j) which appeared in the same session, in response to the same customer query, where one was clicked and the other was purchased. Pairs constructed in this way may contain products that are complementary rather than similar, for example a cell phone and cell phone case.

3.1.2 Data Filtering. Products in our dataset are organized in a category hierarchy. For example, headphones are associated with a path in the category tree: "Electronics"→"Headphone"→...→"Over-Ear Headphones", where "Electronics" is the root category and "Over-Ear Headphones" is the leaf category. To ensure product comparability, we filtered the candidate pairs by keeping only those where both products belong to the same leaf category.

3.1.3 Data Sampling. A product in our dataset might have several variants, such as a different color or size. For such products, we selected one variant and kept only pairs with that variant. To study the relationship between customers' purchase preferences and product quality, we sampled pairs with differing *purchase preferences*. Formally, the purchase preference Z_{ij} for a product pair (i, j), is defined as

$$Z_{ij} = \frac{s_{ij}}{(s_{ij} + s_{ji})} \tag{1}$$

where s_{ij} denotes the number of shopping sessions where i is purchased and j is clicked, and s_{ji} denotes the number of shopping sessions where j is purchased and i is clicked. We binned the pairs according to purchase preference in increments of 0.1. For each bin, we sampled pairs (i, j) proportional to $\log(s_{ij} + s_{ji})$. As a result, we had 946 product pairs summarized in Table 1.

Table 1: Comparable product pairs summary. Note that we count the pair $(i,)$ and (j, i) only once since the order does not make a difference in our user study.

Category	Product Pairs	Total Products
Kitchen	199	308
Electronics	134	194
Home	193	312
Beauty	203	321
Office Supplies	217	349
Total	946	1,484

Table 2: Example of extracted product attributes for product "Bose QuietComfort 35 II".

Data Source	Extracted Attributes
Product Catalog	brand name, wireless communication technology, headphones form factor, microphone form factor, headphones jack, number of boxes, model year, speaker type
Customer-generated content	background noise, noise cancellation, google assistant app, firmware updates, anc issue, high anc, short distance

3.2 Product Attributes Extraction

We extracted attributes from the product catalog and from customer-generated content including customer reviews and Q&A. Table 2 shows an example of the extracted attributes. Note that price is not seen as an attribute since it changes frequently over time.

3.2.1 Catalog Attributes. Attributes in the product catalog are manually created by sellers. While sellers might create a long list of attributes for a product, some will be unimportant to the customer. Similarly, they may also miss important attributes that are represented by similar products. Therefore, we ranked the attributes for every product in the catalog at the same leaf category by their frequency among the products. Then for each product pair we selected the top 10 attributes. If any attributes from the top 10 leaf-level list were missing, they were added to the set. For example, suppose that product A had attributes 3, 4, 6, 7, and 15 and product B had attributes 2, 4, 5, 6, and 8, the aggregate of the two would be products 2, 3, 4, 5, 6, 7, 8, and 15. To this set, we add attributes 1 and 10 from the leaf-level list for a total of 10 attributes.

3.2.2 Attributes of Customer-generated Content. We extracted attributes from customer reviews and product questions and answers. We rely on a standard keyphrase extraction algorithm TextRank [16], which is a graph-based model that ranks noun phrases in a piece of text. Using the implementation from Textacy,[3] we extracted keyphrases from the top five reviews and top three Q&A for the products in our sample.[4] We removed attributes that appear to be opinion words (e.g., bad, good, great, etc.) or local stopwords (such as the product name or category) that remain in the result. In addition, we filtered out attributes based on a manually created list of non-attribute terms (such as "great quality", "good quality", etc).

3.2.3 Attribute Ranking and Selection. After extracting the attributes, we ranked them using a relevance score defined as the linear combination of two language models θ_R and θ_{QA} estimated from the set of attributes derived from customer reviews and product Q&A for all products in the same leaf category. For the reviews R from products in a same leaf category, the probability of a given attribute a is estimated as:

$$P(a|\theta_R) = \frac{c(a,R) + \mu P(a|C)}{|R| + \mu} \tag{2}$$

where $c(a,R)$ is the count of term a in the distribution of terms in R, $|R|$ is the number of terms in the R, $P(a|C)$ is the probability of a given the customer-generated content for all products in the data, and μ is a free parameter. We set $\mu = 1000$, the default setting from Textacy. An analogous language model is estimated from the product Q&A. The score, $P_{R,QA}$, for a term a is the linear combination of the probability that a was generated from a model of the reviews, and a model of the product Q&A:

$$P_{R,QA}(a) = \lambda P(a|\theta_R) + (1-\lambda)P(a|\theta_{QA}) \tag{3}$$

where λ is a parameter between 0 and 1. In our experiment, we set $\lambda = 0.7$ to account for the size of reviews and Q&A for keyphrases extraction. Due to the lack of ground truth for attribute ranking, we manually examined the ranking results, which appear to be reasonable (Table 2). As with catalog attributes, we aggregated the top 10 attributes for both products in the pair, and adding any missing top-10 leaf-level attributes.

4 EXPERIMENTAL SETUP

We recruited crowd workers to compare the quality of product pairs, and to identify the attributes that contributed to their judgment. To bring context to the workers, we asked the workers to play the role of online shoppers by showing them the message in Figure 1 when they landed on the task page.

The workflow of the user study task is shown in Figure 2. Workers start by filling out a before-task questionnaire to provide information about their previous online shopping experience, including how often they shop online and which product category they bought the most, etc. followed by an interactive tutorial on how to use the interface. Product pairs were assigned to workers according to the categories workers indicated they shopped for in the past. After finishing the assessment, workers were presented with an after-task questionnaire about their experience in task execution,

[3]https://github.com/chartbeat-labs/textacy visited October 2019

[4]Reviews and Q&A were ranked using a proprietary algorithm which corresponds to a combination of the usefulness of the review and trust in the reviewer.

> Imagine that you're shopping online for a product and we have recommended you two products that you may want to buy. We have provided you with some information about the products, customer reviews and customer questions & answers. We would like to know some information about these two products, e.g., whether they are similar, which is better, and which attributes do you think are important in assessing their quality. We have matched product pairs that you may find interesting given your profile. After completing the study we will ask you a few more questions about your overall experience.

Figure 1: Message shown to the worker before the task.

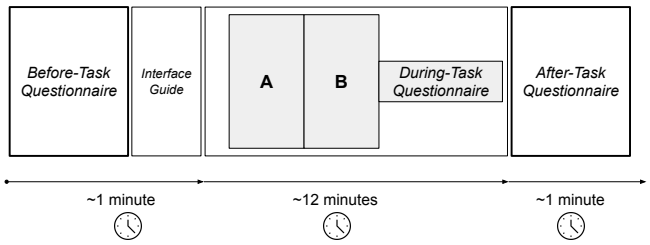


Figure 2: Task workflow.

such as, how easy it was and which data sources were more informative. Details of the before- and after-task questionnaire are shown in Table 3. The overall process took approximately 14 minutes, of which the quality assessment took 12 minutes.

The side-by-side interface is shown in Figure 3, with the user questionnaire on the right side. We randomized the order of products within the pair to avoid presentation bias. We asked workers to use only the information that we provided for their assessments, where the information provided was from the product catalog, the top reviews and product Q&A, and the derived attributes described in Section 3. Workers were asked to indicate whether two products were similar, whether they bought the product in the past, which product was higher quality, and which attributes contributed to their judgments. The order of product attributes was randomized to account for position bias. At the end of the questionnaire, workers were given an opportunity to enter additional attributes not listed. In addition to the explicit information, we collected behavioral signals of workers such as their clicks, scrolls during their task execution and the time taken to complete each task.

We propose that intuitively workers who purchase more items of a particular category are more qualified to assess the quality of products in that category. We assigned product pairs according to the categories indicated by the worker in the pre-task survey. When no such pairs were available, the system then rolled up to the root category or randomly from the remaining product pairs.

Each of the 946 product pairs was assessed by 10 workers, and in each task a worker assessed five pairs. Given the amount of time the task takes, we paid for each task 0.5 USD and a bonus of 1.50 USD if the worker passed the quality check. For each task, we composed the five product pairs such that four of them contained similar products (Section 3.1) and the remaining one contained two obviously dissimilar products. We discarded the results of tasks for which the worker did not correctly identify dissimilar pairs. We

Table 3: Before-task, during-task and after-task questions. During-task questions are asked for each product pair and are presented in the quality assessment interface shown in Figure 3.

Before-Task Questionnaire	
Online Shopping Experience	How often do you buy products online?
Costumer Expertise	Select a product category that you often buy online
Costumer Expertise	Which of these sub-categories in the [chosen category] do you shop online?
Task Affinity	Have you used any website to compare products when you want to buy them online?
During-Task Questionnaire	
Quality control check (F)	Are these products similar?
Customer expertise (G)	Have you bought similar products to these before?
Product quality (H)	Which of the products is of better quality?
Product quality attributes (I)	Which attributes tell you it's better quality?
Product quality attributes (J)	Any other attributes?
Comments (K)	Any additional comments?
Confidence (L)	How confident are you about your judgement?
After-Task Questionnaire	
Similarity	It was easy to decide if two products were similar.
Quality comparison	It was easy to decide which of the products is of better quality.
Quality attributes	It was easy to decide which attributes tell a product is of better quality.
Source usefulness	Did you find the information given to you useful to decide which product was of better quality?

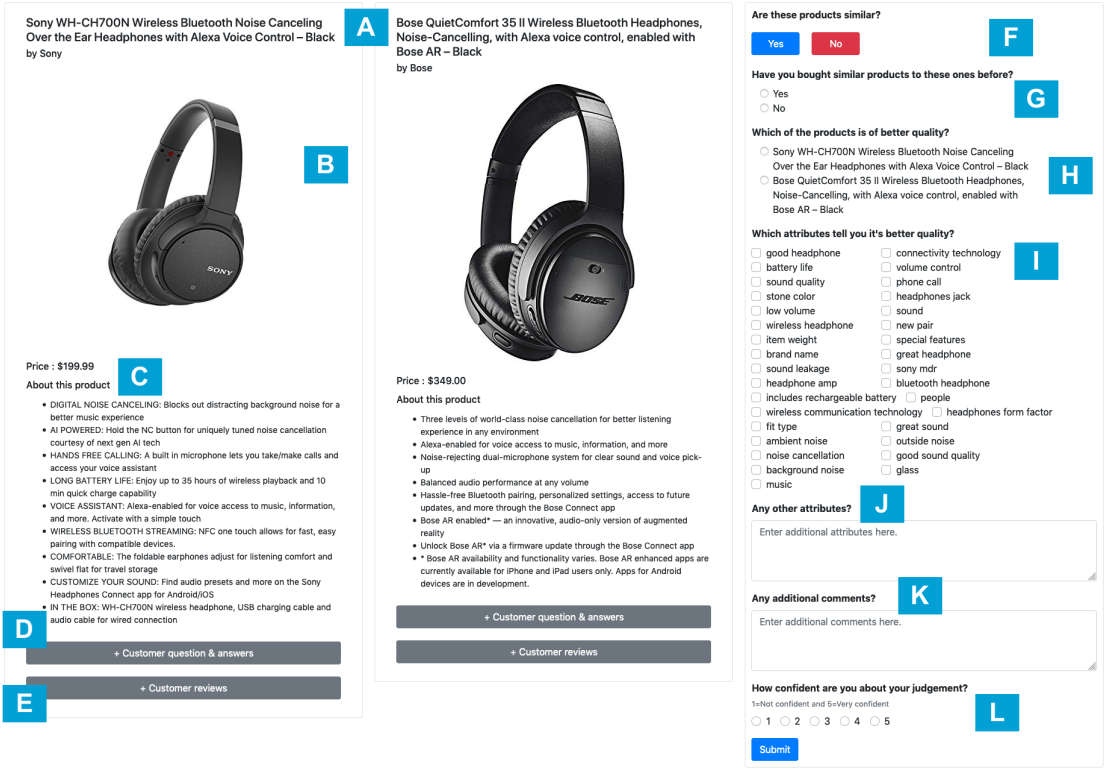
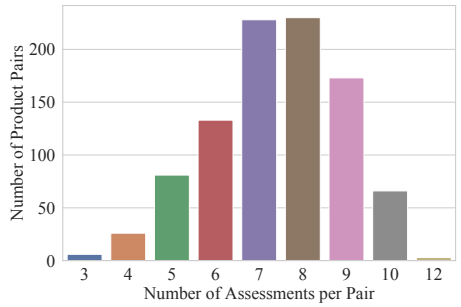
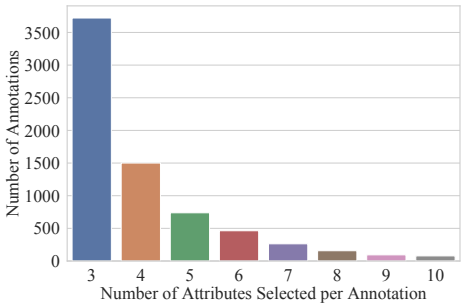


Figure 3: Overview of our interface. Visible are product [A] title, [B] image , [C] price and description, [D] customer question & answers, [E] customer reviews. Remaining parts of during-task questionnaire are described in Table 3.

further filtered out the results from tasks for which the time taken was outside the range $\mu \pm 3 * \sigma$ where μ is the average time took for tasks and σ is the standard deviation.



(a) Number of quality votes per pair



(b) Number of selected attibutes per pair

Figure 4: Descriptive statistics of the assessment: (a) number of quality votes per pair and (b) number of selected attributes per pair.

5 RESULTS

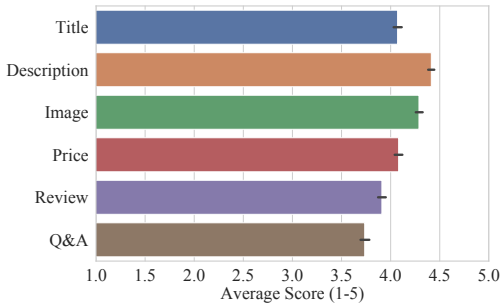
Over the course of five days, we recruited in a total of 420 workers via Figure Eight.[5] Most of our participants came from USA (80%) and the rest from India (11%), UK (5%), Canada (2%), Germany (1.5%), and South Africa (0.5%). Overall, we had 3320 tasks executed (a worker can do more than one task); among them, 2421 tasks were successfully completed. After filtering the results for task execution time, we obtained 8074 annotations (16.5% of annotations were removed).

Figure 4(a) shows the distribution of quality votes. Most pairs received 7-8 quality votes (7.7 on average). Figure 4(b) shows the distribution of selected attributes for per product pair. Workers tended to select a small number of attributes for most quality assessments.

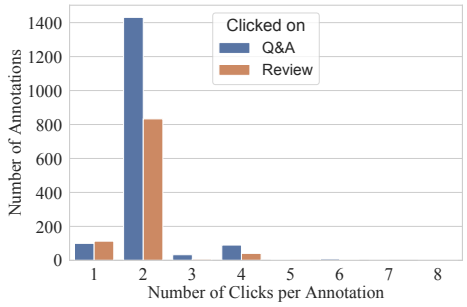
5.1 Which Data Source is More Useful?

To understand whether catalog information or customer-generated information is more useful for quality assessment, we analyzed both explicit feedback that workers provided in the after-task questionnaire and implicit feedback collected from workers' behavioral

[5]https://www.figure-eight.com/



(a) Rated by workers



(b) Review and Q&A clicks

Figure 5: Usefulness of data sources (a) rated by workers and (b) indicated by clicks per annotation.

signals during task execution, as well as the sources of the attributes they selected.

In the after-task questionnaire, workers were asked to evaluate the usefulness of information for quality assessment (Table 3 last row). We asked workers to rate the usefulness of title, description, image, price, reviews and customer Q&As on a scale of one to five. Results are shown in Figure 5(a). We observe that product description and image are deemed the most important, with average scores of 4.4 and 4.3, respectively. Customer-generated content, including reviews and Q&As, is scored least useful. These results correspond to the observed number clicks on the reviews and Q&A per annotation: only 24% of annotations have clicks on the reviews and Q&A. The distribution of clicks per annotation is shown in Figure 5(b). We notice that the number of clicks for Q&A is higher than for reviews despite that Q&A is rated slightly lower. This may be due to position bias as Q&A is always placed above reviews in the task.

A more direct signal for the data source's usefulness is the type of attributes selected by workers. In Figure 6, we show the proportion of selected attributes derived from customer-generated content and catalog. We observe that attributes from review and customer Q&A are consistently selected more often than catalog attributes across different product categories, by a large margin. Such a result suggests that attributes from customer-generated content are in fact more useful for quality assessment, in terms of the number of selected attributes. We note the result is not simply because more attributes were extracted from customer-generated content (see

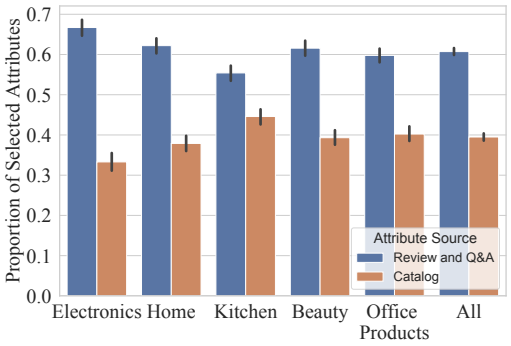
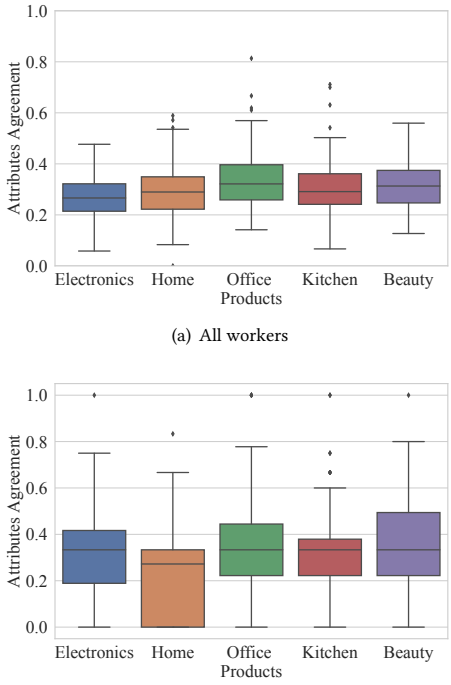


Figure 6: Proportion of selected attributes from catalog and customer-generated content.



(a) All workers

(b) Frequent buyers

Figure 7: Agreement on attributes among (a) all workers and among (b) frequent buyers.

next subsection) since similar numbers of catalog attributes and attributes from customer-generated content were presented for each product pair in the assessment.

We break down the results further into product-level and leaf-category-level attributes, and observe that among catalog attributes, product-level attributes are selected more often than catalog leaf-category-level attributes; whereas for attributes from customer-generated content, we observe that leaf-category-level attributes are selected more often. This is likely due to the difference of the specificity of catalog attributes and attributes derived from customer-generated content. Compared with catalog attributes, attributes derived from customer-generated content are typically much more diverse and product-specific (see Table 4 and more discussion in next subsection). Aggregating attributes derived from customer-generated content on leaf level reduces noise in the attributes and scopes them down to those that are more useful for product comparisons.

The discrepancy between explicit and implicit worker feedback on the usefulness of data sources is likely due to two reasons. First, customer reviews and Q&A take more time to read; workers tend to skip reading such content, as shown by the small number of clicks on reviews and Q&A. Second, the attributes extracted from customer-generated content may be attributes that are commonly known rather than tied to a specific review or Q&A; workers may find it sufficient to read the list of extracted attributes without reading the reviews or Q&A. This corresponds to the observation that the proportion of selected attributes derived from customer-generated content exhibits an minimal difference between annotations regardless of whether the worker clicked on customer-generated content.

5.2 What Attributes Do Customers Care About?

To understand whether there is a consensus about which attributes are important, we first investigate the agreement of attributes selected in quality assessment. We then extract the most selected attributes for each category and analyze the potential difference of important attributes across product categories. We further inspect the manually entered attributes.

We compare the agreement of selected attributes between all annotators and between frequent buyers. Agreement is measured using overlap coefficient, defined as

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}, \tag{4}$$

where X, Y are the sets of attributes selected in two annotations. We computed the average agreement of all pairs of annotations on the same product pair. Figure 7(a) and 7(b) show the agreement of selected attributes for all annotators and frequent buyers, respectively. We observe that frequent buyers demonstrate a higher level of agreement on Electronics and Beauty (p-value $< .01$, Kruskal-Wallis H-test), suggesting that important attributes of products in these categories are more consistently perceived by frequent buyers.

We show in Table 4 the most frequently selected attributes for the five product categories. For each category, we also pick two representative sub-categories and show the most frequently selected attributes in the same table.

We observe that catalog attributes dominate the list of most frequently selected attributes across different product categories. For example, brand name is the most popular attributes across categories, and attributes such as item material and the number of items (i.e., pack size) are always among the top ten. This is likely due to the fact that catalog attributes are typically general attributes that are applicable to a wide range of products, thus selected more often when results are shown on an aggregated level.

Table 4: Top-10 most frequently selected attributes for the five root product categories and the top-5 attributes for two representative sub-categories of these five root categories. Attributes from customer-generated content are underlined.

Electronics	**Kitchen**	**Beauty**	**Home**	**Office Products**
brand name	brand name	brand name	brand name	brand name
sound quality	material	ingredients	material	number of items
great sound	number of pieces	number of items	item weight	item weight
hardware material	number of items	item weight	number of items	connectivity technology
number of items	stainless steel	target gender	is stain resistant	material
connectivity technology	color	long hair	face style	color
color	item weight	material	color	face style
item weight	wattage	color	fabric type	print quality
good sound	dishwasher safe	item volume	finish type	paper size
battery life	capacity	liquid volume	number of pieces	model year
Headphones & Audio	**Cookware, Bakeware**	**Skin Care, Cosmetics**	**Bedding, Bath**	**Office Organization, Office Essentials**
sound quality	cast iron	brand name	fabric type	brand name
great sound	material	ingredients	brand name	number of items
brand name	number of pieces	skin type	material	item weight
good sound	high heat	sensitive skin	is stain resistant	material
connectivity technology	number of items	item volume	item weight	face style
Power, Cables, Other Accessories	**Cooking Appliances, Specialty Electrics**	**Hair Care**	**Home Décor, Arts, Crafts and Sewing**	**Printers**
brand name	brand name	brand name	brand name	brand name
material	material	number of items	number of items	connectivity technology
number of items	stainless steel	hardware material	material	printer technology
power cord	capacity	ingredients	color	good printer
coil voltage	wattage	fine hair	finish type	print quality

Looking at the results on a finer sub-category level, we observe that attributes derived from customer-generated content appear more in the top five. These attributes are more specific, e.g., sound quality is the most frequently selected attribute for Headphones and Audio, and cast iron is the most selected attribute for Cookware and Bakeware. Comparing the overall number of selected attributes, we found that 171 catalog attributes were selected (out of 211) and 2552 attributes from customer-generated content were selected (out of 3312). Such a result shows the diversity of attributes from customer-generated content.

In addition to the attributes from catalog and customer-generated content, annotators were also allowed to enter attributes they found important yet not covered by the list of attributes we provided. We find that 74.2.% of pairs have at least one annotation with manually entered attributes, showing a high-level of worker engagement in the task, however the most commonly entered attribute is "it is not necessary" (13.6% annotations), which we interpret as no extra attributes are used by the workers for quality assessment. Apart from that, the top attribute is "price" (7.3% annotations), which was excluded from the attribute list although it is an important factor in product comparison. Workers also used the textbox to explain the rationale of their selection. For instance, we find feedback such as "The quality is likely the same - I just prefer the multi-colored set, and that attribute isn't listed" and "These are identical apart from the ink color and the price. I chose the first product because of the reviews".

5.3 Is there a Proxy for Quality?

We analyzed potential proxies of quality in the catalog data and customer behavioral data, in particular, customer purchases. We analyze the correlation of perceived product quality to the set of features extracted from catalog and customer behavioral data. To further understand the predictive power of different features for quality perception, we build a random forest regression model and analyze the contribution of each feature to the prediction.

We first define the relative perceived quality of a pair of products (i, j) as the percentage of annotations that prefer product i over j. For each feature in consideration (including customer purchases), we compute a relative value for the feature as

$$f(i,j) = \frac{q(i)}{q(i) + q(j)}, \tag{5}$$

where $q(\cdot)$ represents the corresponding value of the feature for a product in comparison. Note that features we consider all have positive values; consequently, the relative value is a number between 0 and 1. We compute the Spearman's correlation of the features with relative perceived quality and results are shown in Figure 8.

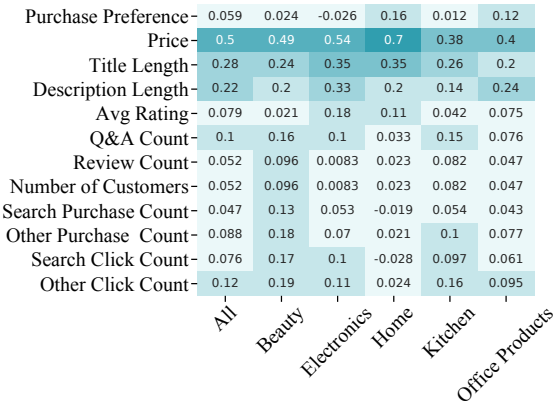


Figure 8: Spearman's correlation between relative product quality and product features from catalog and customer behavioral data.

We observe that price, title length and description length are the three features that significantly correlate with relative perceived quality. These results imply that products with a higher price, longer title and description are generally considered better products. Interestingly, purchase preference does not show a statistically significant correlation. This suggests that quality is not the only consideration in customers' purchase decisions. Considering the fact that products with a higher price are perceived better quality, the result suggests that customers tend to seek a trade-off between price and quality in making purchase decisions.

We further observe that the average rating of a product also does not correlate with the relative perceived quality, either. This could be due to the fact that the average rating was not provided to annotators in quality assessment. Star rating is a complex signal that involves factors beyond the perceived product quality.

Correlation only shows the linear relationship between features and the perceived quality. More complex relationships can be identified by building a nonlinear model using features to predict the perceived quality. We build a random forest regression model [1] to predict the relative perceived quality of each product pair. The model has been proven to be highly effective for a wide range of tasks [5]. We use nested 5-fold cross-validation to search for the hyperparamters and evaluate the performance [3]. The best achieved hyperparameters are 500 trees with a maximum depth of 10. The prediction error is 0.17, which is reasonably small compared with the relative perceived quality that ranges from 0 to 1 (mean=0.5). Using the same hyperparameters, we then retrain the model using the full dataset to estimate the contribution of each feature to the prediction.

Results are shown in Figure 9, where feature importance is defined as the average total decrease in node impurity as contributed by each feature towards the classification result. We observe that the most important features remain to be price, title length, description length. Besides, we also observe that average rating and purchase preference come after them, showing better predictive power than the number of reviews or clicks, which are more related to product popularity than quality.

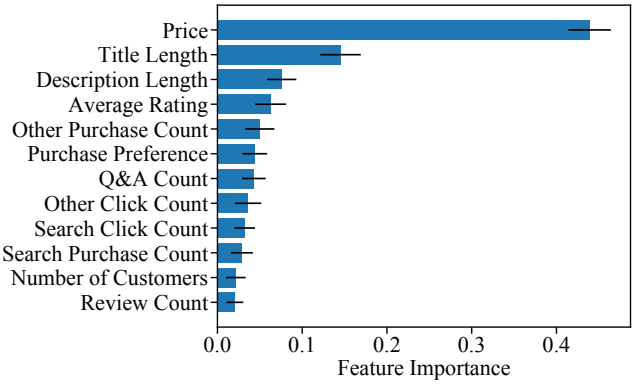


Figure 9: Feature importance for relative perceived quality.

6 CONCLUSIONS AND FUTURE WORK

Implications of the Results. Results show that compared with catalog attributes, attributes from customer-generated content are richer and more diverse, and are more selected in quality assessment. Despite that, customers tend to not read the detailed reviews or Q&A of other customers when they have access to the list of attributes extracted from such content. Another related finding is that for each product pair, customers only rely on a small number of attributes for quality assessments. These results point to the necessity of future research on mining key aspects from customer-generated content to simplify customers' purchase decisions.

We also find that price is strongly correlated with perceived product quality, which confirms previous results from offline studies. More importantly, we show that customers' purchase decisions and ratings are not strongly correlated with product quality. These findings are important for owners of online shopping platforms as it implies that designing non-personalized, quality-based recommenders involves more factors than customers' purchase or rating. The findings are also important for researchers in the field addressing quality-based recommender systems and voice recommenders: there needs to be a shift of focus from using ratings or purchases as the grounding input signal, to using quality-related attributes.

Threats to Validity. We consider the validity threat related to the history effect of the selected product pairs and the selection effect of the product pairs and crowdworkers. The history effect is addressed by collecting product pairs from shopping sessions that span a long and recent observational period which does not contain major public holidays (e.g., Christmas and New Year). The selection effect of product pairs is addressed by sampling from major product categories with consideration of customers' purchase preferences. The size of the product pairs (946) was determined mainly considering the relatively high cost in the crowdsourcing task due to the number of annotations per pair (10) and the expense of each task (2 USD). We note such a setting is important for quality control and payment fairness. We acknowledge that the size of the product pairs and accordingly the number of workers could be limited, but we believe that it does not affect the importance of the results herein presented, but leaves space for future work.

REFERENCES

[1] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[2] Ben Carterette, Paul N Bennett, David Maxwell Chickering, and Susan T Dumais. 2008. Here or there. In *European Conference on Information Retrieval*. Springer, 16–27.

[3] Gavin C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research* 11, Jul (2010), 2079–2107.

[4] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2017. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. *arXiv preprint arXiv:1710.11214* (2017).

[5] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 15, 1 (2014), 3133–3181.

[6] Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: improving rating predictions using review text content. In *Twelfth International Workshop on the Web and Databases*, Vol. 9. 1–6.

[7] Paolo Gastaldo and Judith A Redi. 2012. Machine learning solutions for objective visual quality assessment. In *6th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Vol. 12.

[8] Video Quality Experts Group et al. 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment. In *Video Quality Experts Group Meeting, Ottawa, Canada, March, 2000.*

[9] Tobias Hoβfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not enough!. In *2011 Third International Workshop on Quality of Multimedia Experience*. IEEE, 131–136.

[10] Tobias Hoßfeld, Michael Seufert, Matthias Hirth, Thomas Zinner, Phuoc Tran-Gia, and Raimund Schatz. [n.d.]. Quantification of YouTube QoE via crowdsourcing.

[11] Jacob Jacoby, Jerry C Olson, and Rafael A Haddock. 1971. Price, brand name, and product composition characteristics as determinants of perceived quality. *Journal of Applied Psychology* 55, 6 (1971), 570.

[12] Joseph M Juran et al. 1988. *Juran on planning for quality.* Collier Macmillan.

[13] P Le Callet, C Viard-Gaudin, and D Barba. 2006. A Convolutional Neural Network Approach for Objective Video Quality Assessment. *IEEE Transactions on Neural Networks* 17, 5 (2006), 1316–1327.

[14] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

[15] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.

[16] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. 404–411.

[17] Anush Krishna Moorthy and Alan Conrad Bovik. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing* 20, 12 (2011), 3350–3364.

[18] Vivek Nanda. 2016. *Quality management system handbook for product development companies*. CRC press.

[19] Manish Narwaria and Weisi Lin. 2010. Objective image quality assessment based on support vector regression. *IEEE Transactions on Neural Networks* 21, 3 (2010), 515–519.

[20] Richard W Olshavsky and John A Miller. 1972. Consumer expectations, product performance, and perceived product quality. *Journal of Marketing Research* 9, 1 (1972), 19–21.

[21] Jerry C Olson and Jacob Jacoby. 1972. Cue utilization in the quality perception process. *ACR Special Volumes* (1972).

[22] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 913–921.

[23] Akshay R Rao and Kent B Monroe. 1989. The effect of price, brand name, and store name on buyers' perceptions of product quality: An integrative review. *Journal of Marketing Research* 26, 3 (1989), 351–357.

[24] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).

[25] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 297–305.

[26] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2014. A survey on quality of experience of HTTP adaptive streaming. *IEEE Communications Surveys & Tutorials* 17, 1 (2014), 469–492.

[27] Eugene F Stone-Romero, Dianna L Stone, and Dhruv Grewal. 1997. Development of a multidimensional measure of perceived product quality. *Journal of Quality Management* 2, 1 (1997), 87–111.

[28] Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. 2015. User modeling with neural network for review rating prediction. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[29] Christoph Trattner and Dietmar Jannach. 2019. Learning to recommend similar items from human judgments. *User Modeling and User-Adapted Interaction* (2019), 1–49.

[30] Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Mix'n Match: Integrating Text Matching and Product Substitutability within Product Search. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 1373–1382.

[31] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).