Helping Voice Shoppers Make Purchase Decisions

Gustavo Penha*
g.penha-1@tudelft.nl
Delft University of Technology
Delft, Netherlands

Eyal Krikon
krikon@amazon.com
Amazon
Seattle, USA

Vanessa Murdock
vmurdock@amazon.com
Amazon
Seattle, USA

Sandeep Avula
sandeavu@amazon.com
Amazon
Seattle, USA

ABSTRACT

Online shoppers have a lot of information at their disposal when making a purchase decision. They can look at images of the product, read reviews, make comparisons with other products, do research online, read expert reviews, and more. Voice shopping (purchasing items via a Voice assistant such as Amazon Alexa or Google Assistant) is different. Voice introduces novel challenges as the communication channel is limited in terms of the amount of information people can and are willing to absorb. Because of this, the system should choose the single most effective nugget of information to help the customer, and present the information succinctly. In this paper we report on a within-subject user study ($N = 24$), in which we employed three template-based methods that use information from customer reviews, product attributes and search relevance signals to generate helpful supporting information. Our results suggest that: (1) supporting information from customer *reviews* significantly improves participants perception of system *effectiveness* (helping them make good decisions); (2) supporting information based on search *relevance signals* improves user perception of system *transparency* (providing insight into how the system works). We discuss the implications of our findings for providing supporting information for customers shopping by Voice.

ACM Reference Format:
Gustavo Penha, Eyal Krikon, Vanessa Murdock, and Sandeep Avula. 2022. Helping Voice Shoppers Make Purchase Decisions. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3491101.3519828

1 INTRODUCTION

Voice assistants such as Amazon Alexa and Google Assistant have become widely used in recent years. These devices allow people to interact in their own language to do everyday tasks such as set the kitchen timer, search for a recipe, answer trivia questions, or play music. In addition, these devices allow customers to shop by Voice.

While online shopping on the Web is well-established, people are less inclined to make purchase decisions with a Voice device. On the Web, customers can determine whether a product suits their needs by browsing multiple items, reading reviews, or comparing multiple products. By contrast, Voice customers typically do not go beyond two product recommendations, are more inclined to purchase less expensive products, and do fewer navigational actions [14].

The search and recommender systems which power Voice assistants have invested in improving customer trust by providing supporting information in the form of model explanations [35]. However, directly adopting such techniques to the Voice channel presents a challenge. The real-estate to provide information on a Voice channel is limited and system designers must be judicious in presenting information. More specifically, they must choose the most succinct nugget of information that will help the customer's purchase decision, whether that means providing additional information about the product, or providing transparency into the recommender system itself. Our research adds to the growing body of work on model explanations in search and recommender systems, by investigating how supporting information based on different information sources related to the product can help customers when shopping by Voice.

In this paper, we examine whether supporting information derived from customer reviews, product attributes, or relevance signals from the recommender system is more valuable for customers in making a purchase decision. Building on Balog and Radlinski [1], we propose that in an e-commerce setting, providing supporting information over a Voice channel has four explanation goals: *Persuasiveness* (makes me want to buy a product), *Effectiveness* (provides support for good decision making), *Transparency* (provides insight into how the system works) and *Scrutability* (provides the opportunity for feedback about what the system understood).

To compare these strategies, we created backstories that describe the shopping needs of hypothetical customers. Each backstory represents a specific search intent (the customer has a narrow target or a broad need), and a product consideration (high consideration, for which the customer needs significant additional supporting information to make a purchase decision vs. low consideration, for which the customer needs little additional information to make a purchase decision). For each backstory, we also vary the relevance of the product recommendation by testing for both relevant and non-relevant product recommendation. We included the relevance variation as work by Carmel et al. [4] suggests that e-commerce

*Work done during an internship at Amazon.

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3519828

customers sometimes engage with non-relevant results. We investigated the following research questions:
RQ1: Which source of supporting information (reviews, attributes, or relevance) is most helpful with respect to the explanation goals: persuasiveness, effectiveness, transparency, and scrutability?
RQ2: What is the effect of the search intent (narrow or broad), product consideration (high or low) and the product relevance (relevant or non-relevant) with respect to the explanation goals?

In the remainder of the paper, we present previous work on product search and recommender explainability in Section 2. In section 3, we present details about the user study, including the methods for generating supporting information. In Section 4, we present the results, and discuss them in Section 5.

2 RELATED WORK

Prior work in explanations has taken two approaches to convey information about the underlying systems [37]. In one, the system provides *justifications*, which is information intrinsic to the model. In the second approach, the system provides *descriptions*, which is information acquired from various sources other than the model itself. In this work, we focus on explanation *descriptions*, and investigate the impact of different sources to provide explanations.

Methods for generating explanation *descriptions* in recommender systems can be broadly categorized into two approaches: generation-based methods [6, 7, 12, 19–21, 39] and template-based methods [3, 13, 25, 34, 35]. Generation-based methods use a generative neural network to output a review-like explanation sentence, along with the recommendation score for a $\langle user, item \rangle$ tuple. Template-based methods fill a template with phrases based on item attributes, tags, collaborative information (item and user neighbors) and user demographics. In this paper, as the focus is the choice of information source to provide an explanation, we adopt the template-based method. Through the template based method, we provide supporting information to suggest why the system considered the product to be relevant (such as "the product is frequently shown when people search for...").

The behavior of customers when searching for products in e-Commerce stores is different from general Web search [29]. Purchase decisions are a highly individual process where different factors come into play such as trust in the e-commerce store, price sensitivity, brand affinity, customer reviews, shipping time, and more [16, 27]. Unlike Web search where relevance is generally modeled through lexical and semantic matches between the query and document [22], in product search systems it is necessary to take into account additional factors such as price, ratings, brand reputation, shipping speed, etc. In this work, we examine two types of product information commonly considered by online shoppers: price and customer reviews.

Search intent differs between Product and Web search [28, 32]. Su et al. [33] proposed a taxonomy of customer intents when searching for products: *target finding* (the customer has a specific target in mind with an immediate purchase need), *decision making* (the customer has a vague idea of what to buy but would like to explore and compare products) and *exploration* (the customer has no specific target in mind and is browsing). In this work, we compare explanation strategies for target finding and decision making.

Tintarev and Masthoff [35] describe different goals of recommender explanations: *transparency, scrutability, trust, effectiveness, persuasiveness, efficiency* and *satisfaction*. Following Balog and Radlinski [1] we adapt the following explanation goals to the e-commerce setting: *transparency* (provide insight into how the system works), *scrutability* (providing the opportunity for feedback about what the system understood), *effectiveness* (providing support for good decision making) and *persuasiveness* (providing reasons to buy a product). Motivated by differences between Voice and Web shopping [14], we study how these goals are affected by the type of information surfaced in the Voice interaction, given the search intent and the product consideration. Ultimately, we hope to determine the most helpful information type for Voice product search, given the necessary brevity of the Voice interaction.

3 METHODS

We conducted a user study by constructing template responses of a Voice product search system that introduces four different types of supporting information. This was a $4x2^3$ mixed-level factorial design, where each task has a backstory associated with a search intent (narrow or broad) and a product consideration (high consideration or low). In addition, for each of the backstories we presented a relevant product, and a related but non-relevant product, as people are often presented (and engage with) non-relevant products related to their query [4].

3.1 Voice Shopping Templates

For the user study, we created responses with template-based[1] structures [25], that has a baseline response followed by a slot for the supporting information. Following Zhang and Chen [39], the supporting information consisted of product attributes, information from reviews or salient terms important to the ranking model (*relevance words*). Examples of each type of information can be found in Table 1. We aimed for short helpful review sentences and a small number of relevance words to decrease the overhead of listening to a long audio prompt.

The baseline template is similar to typical responses of Voice enabled product search systems. The information about the recommended product for a query includes the product title, the price and the number of days for delivery. Unlike Web search where several attributes from the product can be seen on the product page, Voice is limited to a single additional product attribute.

As shown in Moraes et al. [24], price is a key attribute for customers in estimating product quality, so we included a discount by adding the sentence '*The product is now on sale*'. No discount was given or factored into the price to avoid introducing a confounding factor. When making a purchase decision, customers often turn to reviews from other customers. Following Gamzu et al. [10], we selected a sentence from a positive review[2] of the product with the highest count of helpful votes. Finally, to help the customer understand what makes the suggested product relevant to their need, we manually select words that indicate a match between the product

[1] A generative approach would lead to unnecessary language variations for the purpose of this study.
[2] We do not explore negative reviews as picking a negative review sentence to recommend a product is counterintuitive.

Table 1: Templates for supporting information method, an example prompt for the query '*I want to buy a Fuji camera X100V*' and the average audio length (in seconds). Bold sentences highlight the difference compared to the baseline.

Method	Template	Example	Length (in seconds)
Baseline	`I found` ⟨*product_title*⟩. `It's` ⟨*price*⟩. `With delivery in` ⟨*delivery*⟩ `days`	I found Fujifilm Digital Camera Black. It's 1449.99 dollars, with delivery in 3 days.	11.1
Attributes	`I found` ⟨*product_title*⟩. `The product is now on sale for` ⟨price⟩. `With delivery in` ⟨*delivery*⟩ `days.`	I found Fujifilm Digital Camera Black. **The product is now on sale for** 1449.99 dollars, with delivery in 3 days	12.9
Reviews	⟨*baseline*⟩. `A reviewer said` ⟨*review_sentence*⟩.	I found Fujifilm Digital Camera Black. It's 1449.99 dollars, with delivery in 3 days. **A reviewer said '*The street/documentary/everyday photographer's best tool.*'**.	16.4
Relevance	⟨*baseline*⟩. `The product is frequently shown when people search for` ⟨*relevance_words*⟩.	I found Fujifilm Digital Camera Black. It's 1449.99 dollars, with delivery in 3 days. **The product is frequently shown when people search for '*X100V camera*'.**	15.5

and the query. In a real system, the information presented here might be derived from an automated model explanation [9, 30, 36]. The product attributes such as top reviews and price were extracted from their respective web pages on a large e-commerce website. For product titles, we used shorter voice-friendly versions typical of Voice devices.

3.2 User Study Design

We employ a 4×2^3 mixed-level factorial design with the following independent variables: three two-level factors (the search intent, the product consideration, and the product relevance) and four supporting information methods (the baseline and methods based on reviews, attributes and relevance). The dependent variables are the following explanation goals: *Persuasiveness*, *Effectiveness*, *Transparency* and *Scrutability*. In the following sections we explain the conditions of the user study, the procedure taken by participants and the methodology for analyzing the results.

Each task in the user study has a backstory which describes the underlying information need and product combination. Each backstory is composed of a text to describe the information need and a voice query associated with that need. For examples of the backstories and voice queries we refer the reader to Table 2. Additionally, each backstory is associated with a search intent and product consideration. For each of the four backstories, we have one relevant product and one that is not relevant.

We include two search intents for describing the search behavior that are representative of the way consumers engage with product search systems [33]: *Target-Finding* and *Decision-Making*. With Target-Finding, the customer has a specific item in mind (known product name or brand) with an immediate purchase need, and usually does not compare different products. Whereas with Decision-Making the customer also has an immediate purchase need but compares related products in order to make a purchase decision.

For each intent, we include a *High Consideration* product and a *Low-Consideration* product. High consideration products typically require product research or comparisons and are often more expensive. Low consideration products do not require product research or comparisons, and are typically cheaper. We investigate the impact of product consideration on the different goals.

Product search engines are not always effective, and in some cases non-relevant results are presented to users. For voice product search, this is critical as users typically only hear one or two product recommendations, displaying a bias toward default choices [23]. Interestingly, Carmel et al. [4] found that in some cases users purchase the non-relevant products anyway. Motivated by this finding, we also study the effect of the product relevance.

For the backstories with Target-Finding intent and a non-relevant product, we show a product that matches the product category of the information need, but is not the specific product mentioned in the query. For example, in the backstory B1 the query mentions the camera model *"Fujifilm X100V"*" and the non-relevant system response is a different camera model ("*Ricoh GR III*"). For the backstories with Decision-Making intent and a non-relevant product, we show a product that does not match the product category of the query, but is related to it. For example, in the backstory B3 the query asks for a camera and the non-relevant system response is a camera bag.

Half of the trials for each backstory were populated with relevant results, and the other half with non-relevant results. The order of the trials, supporting information method, and backstory were counter-balanced using balanced Latin squares. Each subject went through the following steps for each of the four methods: (1) read the backstory and query, (2) listen to the audio with the response and (3) answer a questionnaire regarding the different goals. The audios were recorded using the Alexa Presentation Language Audio.[3] The study was conducted through a website without time limitation and participants could hear the explanation multiple times.

In order to evaluate the supporting information we rely on post-task questionnaires, which is the most common methodology for

[3]https://developer.amazon.com/en-US/docs/alexa/alexa-presentation-language/apla-document.html visited January 2022

Table 2: Different task conditions used in the study. Each backstory describes the underlying information need, which has search intent (target-finding vs. decision making), product consideration (high-consideration vs low-consideration) and product relevance (relevant vs. non-relevant). The product title is followed by supporting information (review-based in these examples).

Backstory	Search Intent	Product Consideration	Relevant	Non-Relevant
B1 *You've recently watched some videos on youtube about street photography and you found it really fascinating. So you want to start practicing it. You believe that the camera of your smartphone does not fit the requirements for street photography. After watching several videos of photographers comparing different digital camera options for street photography, you are somewhat convinced that the right choice for you is a camera named "Fujifilm x100v", which is compact but powerful. So you decide to buy it through a voice assistant.* (query: "buy Fujifilm X100V")	Target Finding	High Cons	*Fujifilm Digital Camera Black* (X100V camera) [The street / documentary / everyday photographer's best tool.]	Ricoh GR III Digital Compact Camera, 24mp, 28mm F 2.8 Lens (snapshot camera) [It deserves far more attention than has received.]
B2 *You've realized you are out of deodorant. You decide to buy the same product you used last time from the brand Degree Men using a voice assistant.* (query: "buy deodorant degree men")	Target Finding	Low Cons	*Degree Men,Cool Rush Antiperspirant, 6 pack of 2.7 oz each* (degree men deo) [I get complimented on how I smell constantly.]	*Dove Men +Care Deodorant Stick [...]* (men antiperspirant) [The best 'standard' deodorant you can buy.]
B3 *You've recently watched some videos on youtube about street photography and you found it really fascinating. So you want to start practicing it. You believe that the camera of your smartphone does not fit the requirements for street photography. So you decide to look for street photography cameras and make a decision with the help of a voice assistant.* (query: "buy street photography camera")	Decision	High Cons	*Ricoh GR III Digital Compact Camera, 24mp, 28mm F 2.8 Lens* (snapshot camera) [It deserves far more attention than has received.]	*Leather Camera Bag, Street Photography [...]* (street photography camera bag) [It's a beautiful bag, I've gotten many compliments on it]
B4 *You've realized you are out of deodorant. You did not quite like the quality of the last deodorant you bought, so you decide to explore different options using a voice assistant.* (query: "buy deodorant")	Decision	Low Cons	*Dove Men +Care Deodorant Stick [...]* (men antiperspirant) [The best 'standard' deodorant you can buy.]	*Dial Antibacterial Deodorant Bar Soap, 4oz Each, Pack of 3 Gold Bars* (deodorant bar soap) [It smells good, and it cleans my hands well.]

this task [26]. The questions were adapted from [1] to reflect a product search scenario:

- scrutability - *"Does the supporting information provide enough information to allow a critique of the system"*
- transparency- *"Does the supporting information explain how the system works"*
- effectiveness- *"Does the supporting information help the customer make a good decision"*
- persuasiveness- *"Does the supporting information convince the customer to buy"*

Each question was rated on a 4-point Likert scale: (1) not at all, (2) moderately, (3) slightly and (4) a great deal, for each of the goals.

A total of 24 participants took part in the user study, resulting in a total of 96 distinct evaluations of supporting information. Table 4 shows the average participant scores for each method and factor combination. All participants are researchers from a large tech company and participated in the study remotely.

We use linear mixed-effect models [2], using the `lme4` library[4] to evaluate the effect of different conditions on the supporting information methods. The β coefficients are reported in Table 3. Linear mixed-effects models are better suited for these evaluations over ANOVA as they enable us to take into account multiple fixed effects, their interactions, and random effects and to handle imbalanced data [11]. Our fixed effects are the supporting information methods, the search intent, the product consideration and the product relevance. We treat the *subject id* as a random effect. To test for the significance of the different factors in the dependent variables (the explanation goals) we use the Likelihood Chi-Square Ratio Test [17] which compares the full model with a model without the fixed factor for supporting information.

When testing the interactions between the factors and the proposed explanation methods, we do not simply explore all possible interactions, as it can significantly increase the chance of false

[4]https://github.com/lme4/lme4 visited March 2022

positives. Instead, we formulate three hypotheses based on prior work [4, 15, 33], and only test for them.

H1: *The interaction between the product relevance and relevance-based information is significant.*

Intuitively, a customer might not find relevance words helpful if the product is already relevant for the query. However, as pointed out by Carmel et al. [4], the less obvious cases where the product seems to be non-relevant but is related to the query could benefit from relevance-word information.

H2: *The interaction between the product type (low/high consideration) and review-based information is significant.*

Intuitively, a customer may find review information helpful if a product is more expensive and the purchase decision needs more consideration. For example, it was found that consumers use reviews more in the consideration stage [15].

H3: *The interaction between the search intent and attribute-based information is significant.*

Intuitively, a customer might not worry about the price—the attribute we consider in our experiments—if the intent is target finding where the user is more certain (and knowledgeable) about the product she wants [33].

4 RESULTS

In the following section, we present results from our research questions and hypotheses.

RQ1: Supporting Information Methods In RQ1, we investigated the effect of supporting information methods on four goals. We found that the supporting information methods had a significant effect on two goals: *Effectiveness* ($\chi^2(3) = 8.373, p < 0.05$), and *Transparency* ($\chi^2(3) = 17.352, p < 0.01$). In terms of *Effectiveness*, participants reported that they were better able to determine how well they liked a product using review-based information in comparison to the baseline ($\beta = 0.375, S.E = 0.188, p < 0.05$). In terms of *Transparency*, participants reported that they had a better understanding of the reasoning behind their recommendation from relevance-based keywords in comparison to the baseline ($\beta = 0.973, S.E = 0.184, p < 0.001$).

RQ2: Search Intent, Product Consideration and Relevance In RQ2, we investigated the main effect of three factors on the goals: search intent, product consideration and product relevance. We found that product consideration and product relevance significantly affect various goals and we report on the effects.

Product Consideration: We found product consideration has a significant effect for *Persuasiveness* ($\chi^2(1) = 18.769, p < 0.01$), *Effectiveness* ($\chi^2(1) = 19.725, p < 0.01$) and *Transparency* ($\chi^2(1) = 4.481, p < 0.05$). In terms of *Persuasiveness*, participants reported that they were more likely to be persuaded to purchase low consideration products in comparison to high consideration products ($\beta = 0.730, S.E = 0.165, p < 0.001$). In terms of *Effectiveness*, participants reported that supporting information methods were more helpful when deciding over low-consideration products in comparison to high-consideration products ($\beta = 0.605, S.E = 0.133, p < 0.001$). Finally, in terms of *Transparency*, participants reported that they better understood the rationale for their recommendations for low-consideration products in comparison to high-consideration products ($\beta = 0.380, S.E = 0.184, p < 0.01$).

Product Relevance: We found product relevance to have a significant effect for *Persuasiveness* ($\chi^2(1) = 11.732, p < 0.01$). Unsurprisingly, particpants reported to being persuaded to purchase relevant products in comparison to non-relevant products ($\beta = 0.563, S.E = 0.165, p < 0.001$).

H1-H3: Interaction with Supporting Information methods: In **H1**, we found no significant interaction effects between product relevance and relevance-based keywords across all explanation goals. In **H2**, we found no signficant interaction effects between product consideration and review-based information for *Persuasiveness* ($\chi^2(1) = 2.898, p = 0.088$) . In **H3**, we found significant interaction effects between search intent and attribute-based information for *Persuasiveness* ($\chi^2(1) = 0.017, p < 0.05$) but not significant for *Transparency* ($\chi^2(1) = 3.14, p = 0.076$).

H2: We found trends that suggest review-based information is better suited for persuading participants to purchase high-consideration products in comparison to low-consideration products ($\beta = 0.685, S.E = 0.4, p = 0.089$).

H3: In this analysis, we made two observations. First, it was easier to persuade participants to purchase through attribute-based information for target-finding in comparison to decision-making tasks ($\beta = 0.950, S.E = 0.4, p < 0.05$). Second, we found trends which suggest that attribute-based information was better suited to help participants understand the rationale behind their recommendations for target-finding in comparison to decision-making tasks ($\beta = 0.788, S.E = 0.44, p = 0.075$).

5 DISCUSSION

We found that review-based supporting information significantly helps participants in assessing a product (*Effectiveness*). These findings are aligned with prior work on textual explanations on the Web [5, 18, 38, 40]. Intuitively, this makes sense as reviews contain word-of-mouth information, which prior work suggests helps in decision-making [8, 27]. This motivates further study of models that automatically find short and relevant spans of text from reviews to be used in Voice product search such as Gamzu et al. [10]. As we are studying Voice interactions, we selected review snippets that were coherent and brief (as can be seen from the audio length in Table 1). Automatically identifying helpful snippets from reviews that are also brief and coherent are challenges for future work.

Supporting information in the form of relevant words were useful in improving *Transparency* and *Scrutability*. The words provided evidence for participants to understand the rationale behind the recommendation, and sufficient information to be able to critique the system. For the Voice domain, this could be an effective technique as the information bandwidth is limited and selecting one or two relevant words versus coming up with an entire sentence could be less risky. Future work should continue building on recent efforts such as selecting terms to explain top retrieved products [30, 36] and system interpretation of the query [31], for the Voice domain.

One might expect that if a customer issues a specific search query (a target-finding task) they might find less value in supporting information, as they have already have a clear idea of what they want. We found this to be true for attribute-based information, but not for other types of supporting information. More specifically, we found that when participants received attribute-based information,

Table 3: The β coefficients (with standard deviations) for the fixed effects in a linear mixed-effects regression for the four explanation goals. Asterisks "$*$" and "$**$" denote significance at p<0.05 and p<0.001 respectively

	Persuasiveness	*Effectiveness*	*Transparency*	*Scrutability*
Reviews	0.188 (0.233)	**0.375*** (0.188)	0.411 (0.260)	0.071 (0.185)
Attributes	−0.272 (0.233)	−0.111 (0.188)	0.041 (0.260)	−0.033 (0.184)
Relevance	−0.084 (0.233)	0.224 (0.188)	**0.973**** (0.260)	0.319 (0.185)
Search intent	0.189 (0.165)	0.215 (0.133)	0.032 (0.184)	0.184 (0.131)
Product consideration	**0.730**** (0.165)	**0.605**** (0.133)	**0.380*** (0.184)	0.086 (0.131)
Product relevance	**0.563**** (0.165)	0.022 (0.133)	0.338 (0.184)	0.128 (0.131)

Table 4: Average score for each supporting information method by each of the factors (Search intent, Product consideration, Product relevance). Bold values indicate which group has the highest average for each goal.

Search intent								
	Persuasiveness		*Effectiveness*		*Transparency*		*Scrutability*	
	Decision	Target	Decision	Target	Decision	Target	Decision	Target
Baseline	**2.58**	2.15	**2.42**	2.00	**2.42**	1.77	**1.92**	1.69
Reviews	2.42	**2.77**	2.25	**2.92**	2.25	**2.77**	**2.00**	1.77
Attributes	1.62	**2.58**	1.69	**2.50**	1.77	**2.50**	1.54	**2.00**
Relevance	**2.38**	2.25	**2.54**	2.33	**3.31**	2.83	1.77	**2.50**
Average	2.25	2.44	2.22	2.44	2.44	2.47	1.81	1.99
Product consideration								
	Persuasiveness		*Effectiveness*		*Transparency*		*Scrutability*	
	HighCons	LowCons	HighCons	LowCons	HighCons	LowCons	HighCons	LowCons
Baseline	1.85	**2.92**	1.62	**2.83**	2.00	**2.17**	1.62	**2.00**
Reviews	2.50	**2.69**	2.42	**2.77**	2.17	**2.85**	**2.00**	1.77
Attributes	1.62	**2.58**	1.77	**2.42**	1.92	**2.33**	1.46	**2.08**
Relevance	1.92	**2.69**	2.33	**2.54**	2.92	**3.23**	**2.33**	1.92
Average	1.97	2.72	2.04	2.64	2.25	2.65	1.85	1.94
Product relevance								
	Persuasiveness		*Effectiveness*		*Transparency*		*Scrutability*	
	nonrelevant	Relevant	nonrelevant	Relevant	nonrelevant	Relevant	nonrelevant	Relevant
Baseline	1.92	**2.83**	2.15	**2.25**	2.00	**2.17**	**1.85**	1.75
Reviews	2.33	**2.85**	**2.67**	2.54	2.58	**2.46**	1.75	**2.00**
Attributes	1.69	**2.50**	**2.08**	**2.08**	1.62	**2.67**	1.69	**1.83**
Relevance	2.25	**2.38**	2.33	**2.54**	2.92	**3.23**	2.00	**2.23**
Average	2.05	2.64	2.31	2.35	2.28	2.63	1.82	1.95

they were more easily persuaded and had a better understanding of their recommendation for target-finding tasks. Across all other types of supporting information and goals, we found no difference between target-finding and decision-making tasks. Perhaps for decision-making tasks, participants were interested in attributes beyond price, and therefore did not find the additional information helpful, especially given that the price was already mentioned in the baseline template.

Our results suggest that product consideration had a significant effect for *Persuasiveness*, *Effectiveness*, and *Transparency*. The supporting information for low-consideration products was rated higher than for high-consideration products. Interestingly, review-based information proved to be better at persuading participants to purchase high-consideration products in comparison to low-consideration products. This may be because word-of-mouth plays a greater role in persuading participants, or reviews may mention attributes that are otherwise not present in the product creatives.

As expected, it was easier to persuade participants to purchase relevant products. For the rest of the goals, we did not find any significant differences between relevant and non-relevant products. The lack of significant differences contradicts the hypothesis in Carmel et al. [4] that supporting information is more beneficial for suggesting non-relevant products. We posit that our evidence may

not be conclusive, and that non-relevant products may need specialized supporting information that includes familiarity, product popularity and personalized information [4].

One limitation of the study is that the participants are all researchers from a large tech company, and may not be typical of Online shoppers. Additionally, our study is based on non-interactive and scenario-based audio files as opposed to consumer behavioral data with an actual purchase intent. This along with our unique selection of queries, products, attributes and small number of participants may limit the generalizability of our findings.

6 CONCLUSION

In this paper we examined different sources for supporting information and their utility toward four goals of *Persuasiveness*, *Effectiveness*, *Transparency* and *Scrutability* in Voice product search. We conducted a user study where each task was defined by a backstory associated with a search intent (narrow or broad) and a product consideration (high consideration or low). In addition, for each of the backstories we presented a relevant product, and a related but non-relevant product, as people are often presented (and engage with) non-relevant products related to their query.

We found that supporting information helps in decision-making, and improves the transparency of system recommendations. Review-based information was most helpful with decision making when participants needed to spend more time investigating a product (i.e. high-consideration purchases). Relevance-based information was best suited for improving recommender transparency. Attribute-based information was most useful for *Persuasiveness* and *Transparency* for target-finding tasks. Future work should examine the impact of other product attributes (beyond price) as well as automating the generation of supporting snippets.

REFERENCES

[1] Krisztian Balog and Filip Radlinski. 2020. Measuring recommendation explanation quality: The conflicting goals of explanations. In *Proceedings of the 43rd International ACM Sigir Conference on Research and Development in Information Retrieval*. 329–338.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

[3] Roi Blanco, Diego Ceccarelli, Claudio Lucchese, Raffale Perego, and Fabrizio Silvestri. 2012. You Should Read This! Let Me Explain You Why. (2012).

[4] David Carmel, Elad Haramaty, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek. 2020. Why Do People Buy Seemingly Irrelevant Items in Voice Product Search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM'20). ACM.*

[5] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural A entional Rating Regression with Review-level Explanations.(2018). (2018).

[6] Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. 2021. Generate natural language explanations for recommendation. *arXiv preprint arXiv:2101.03392* (2021).

[7] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-Attentive Multi-Task Learning for Explainable Recommendation.. In *IJCAI*. 2137–2143.

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

[9] Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. 2019. A study on the Interpretability of Neural Retrieval Models using DeepSHAP. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1005–1008.

[10] Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 2021. Identifying Helpful Sentences in Product Reviews. *arXiv preprint arXiv:2104.09792* (2021).

[11] Ralitza Gueorguieva and John H Krystal. 2004. Move over anova: progress in analyzing repeated-measures data andits reflection in papers published in the archives of general psychiatry. *Archives of general psychiatry* 61, 3 (2004), 310–317.

[12] Deepesh V. Hada, Vijaikumar M., and Shirish K. Shevade. 2021. ReXPlug: Explainable Recommendation Using Plug-and-Play Language Model. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 81–91. https://doi.org/10.1145/3404835.3462939

[13] Diana C Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of argumentative explanation types on the perception of review-based recommendations. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 219–225.

[14] Amir Ingber, Arnon Lazerson, Liane Lewin-Eytan, Alexander Libov, and Eliyahu Osherovich. 2018. The challenges of moving from web to voice in product search. In *Proc. 1st International Workshop on Generalization in Information Retrieval (GLARE 2018). http://glare2018. dei. unipd. it/paper/glare2018-paper5. pdf.*

[15] Sungha Jang, Ashutosh Prasad, and Brian T Ratchford. 2012. How consumers use product reviews in the purchase decision process. *Marketing letters* 23, 3 (2012), 825–838.

[16] Teklehaimanot Tadele Kidane and RRK Sharma. 2016. Factors Affecting Consumers' purchasing Decision through ECommerce. In *Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia*, Vol. 8. 159–165.

[17] Gary King. 1989. *Unifying political methodology: The likelihood theory of statistical inference*. Cambridge University Press.

[18] Trung-Hoang Le and Hady W Lauw. 2021. Explainable Recommendation with Comparative Constraints on Product Aspects. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 967–975.

[19] Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 755–764.

[20] Lei Li, Yongfeng Zhang, and Li Chen. 2021. EXTRA: Explanation Ranking Datasets for Explainable Recommendation. *arXiv preprint arXiv:2102.10315* (2021).

[21] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. *arXiv preprint arXiv:2105.11601* (2021).

[22] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond. *arXiv preprint arXiv:2010.06467* (2020).

[23] Alex Mari and René Algesheimer. 2021. The role of trusting beliefs in voice assistants during voice shopping. (2021).

[24] Felipe Moraes, Jie Yang, Rongting Zhang, and Vanessa Murdock. 2020. The role of attributes in product quality comparisons. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 253–262.

[25] Cataldo Musto, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2019. Justifying recommendations through aspect-based sentiment analysis of users reviews. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 4–12.

[26] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3 (2017), 393–444.

[27] Nia Budi Puspitasari, Susatyo Nugroho WP, Deya Nilan Amyhorsea, and Aries Susanty. 2018. Consumer's buying decision-making process in E-commerce. In *E3S Web of Conferences*, Vol. 31. EDP Sciences, 11003.

[28] Nikitha Rao, Chetan Bansal, Subhabrata Mukherjee, and Chandra Maddila. 2020. Product insights: Analyzing product intents in web search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2189–2192.

[29] Jennifer Rowley. 2000. Product search in e-shopping: a review and research propositions. *Journal of consumer marketing* (2000).

[30] Jaspreet Singh and Avishek Anand. 2019. Exs: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 770–773.

[31] Jaspreet Singh and Avishek Anand. 2020. Model agnostic interpretability of rankers via intent modelling. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 618–628.

[32] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A taxonomy of queries for e-commerce search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1245–1248.

[33] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User intent, behaviour, and perceived satisfaction in product search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 547–555.

[34] Panagiotis Symeonidis. 2008. Justified recommendations based on content and rating data. In *WebKDD Workshop on Web Mining and Web Usage Analysis*. ACM.

[35] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[36] Manisha Verma and Debasis Ganguly. 2019. LIRME: locally interpretable ranking model explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1281–1284.

[37] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent*

user interfaces. 47–56.

[38] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 165–174.

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

[40] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.