Data in Brief 54 (2024) 110353

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# Survey data on customer two-stage decision-making process in household vacuum cleaner market

Yinshuang Xiao[a,1], Yaxin Cui[b,1], Nikita Raut[c], Jonathan Januar[d], Johan Koskinen[e], Noshir Contractor[b], Wei Chen[b], Zhenghui Sha[a,*]

[a] The University of Texas at Austin, Austin, TX 78712, United States of America
[b] Northwestern University, Evanston, IL 60208, United States of America
[c] Amazon, Los Angeles, CA 90036, United States of America
[d] The University of Melbourne, Parkville VIC 3010, Australia
[e] Stockholm University, Frescativägen, 114 19 Stockholm, Sweden

ARTICLE INFO

*Article history:*
Received 29 May 2023
Revised 14 March 2024
Accepted 18 March 2024
Available online 26 March 2024

Dataset link: Survey Data on Customer Two-Stage Decision-Making Process in Household Vacuum Cleaner Market (Original data)

*Keywords:*
Customer preference
social influence
consideration-then-choice decision-making
product information retrieval
product design

ABSTRACT

This paper presents the data collection method and introduces the dataset about consumers' consider-then-choose behaviors in the household vacuum cleaner market. First, we designed a questionnaire that collected participants' consideration and choice data, social network data, demographic information, and preferences for product features. In addition, we obtained data on vacuum cleaner product features through web scraping from online shopping websites. After data cleaning and processing, the resulting dataset enables investigation into customer preferences in two stages, namely the consideration and choice stages and the impact of social influence on the two-stage decision-making process. This dataset is unique as it is the first of its kind to collect both customers' revealed preferences in a two-stage decision-making process and their ego social networks. This enables the modeling of customer preferences while accounting for social influence. The published survey questionnaire can be used as a template to collect data on other products in support of customer preferences modeling and the design for market systems.

* Corresponding author.
*E-mail address:* zsha@austin.utexas.edu (Z. Sha).
[1] These two authors contributed equally to this work

https://doi.org/10.1016/j.dib.2024.110353

Specifications Table

| | |
|---|---|
| Subject | Marketing |
| Specific subject area | Vacuum cleaner product attribute data; customer preference data in the household vacuum cleaner market; and customers' ego-centric social network data |
| Type of data | Table (.csv format)<br>Survey questionnaire (.pdf format) |
| Data collection | The vacuum cleaner attribute data were acquired by web crawling the mainstream online shopping platforms in the US market (Amazon, Wayfair, Best Buy, Home Depot, and Walmart) from product specifications and manuals. Subsequently, missing values were filled out, and noisy data were corrected manually by searching product catalogs online.<br>The customers' preference data were collected through a survey questionnaire on a website, both designed by the authors. The survey was launched by the Cint Platform, a digital insight gathering platform with quality assurance mechanisms. The survey was distributed to individuals who had recently purchased a vacuum cleaner and administered over two months, from April 25 to June 25, 2021. In total, 1002 responses were received. This survey was conducted with the approval of the Institutional Review Boards at the University of Arkansas and Northwestern University.<br>Instruments: Cint Platform for launching surveys; Python and Structured Query Language (SQL) for data collection, storage, and query; Microsoft Excel, R, and Python for data cleaning. |
| Data source location | Country: United States |
| Data accessibility | Repository name: Texas Data Repository<br>Data identification number: doi:10.18738/T8/SPJSLI<br>Direct URL to data:<br>https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/SPJSLI<br>(This data is open access with a Creative Commons Attribution (CC-0) license) |
| Related research article | *Xiao, Y., Cui, Y., Raut, N., Januar, J., Koskinen, J., Contractor, N., Chen, W., Sha, Z. (2022). Information Retrieval and Survey Design for Two-Stage Customer Preference Modeling. Proceedings of the Design Society, 2, 811-820.* doi:10.1017/pds.2022.83 [1] |

## 1. Value of the data

- The datasets are useful for customer preference modeling. In particular, this dataset includes customers' ego-centric social network data and their preferred product selections in both the consideration and choice stages. So, the datasets can support the investigation of social influence on customers' choices and their consideration-then-choice modeling.
- The dataset can be used to assess the impact of product attributes (e.g., price, weight, suction power, etc.) and customer attributes (e.g., household size, demographic attributes, personal viewpoints, etc.) on the consideration and selection of products. Besides, the data can be used to study competition among different vacuum cleaner brands and manufacturers.
- The datasets can serve as a means to validate the reproducibility and repeatability of many existing customer preference-related models, which have previously relied on inaccessible commercial datasets.
- In addition to their research applications, the datasets are also suitable for educational purposes in engineering product design and survey methodology. For example, the understanding of customer preferences to product attributes at different stages (consideration and choice) will be important for students to learn the concept of user-centered design.
- The primary beneficiaries of the data include engineering product designers, marketing specialists, and researchers from both engineering and marketing science, as well as digital platform entrepreneurs seeking to develop and refine their products.

- The questionnaire used in this study comprises six sections designed for individuals who are new buyers of household vacuum cleaners. These sections include questions about: 1) the products considered and purchased by customers; 2) the impact of social networks on decision-making; 3) factors influencing decision-making in purchase; 4) personal viewpoint; 5) product usage context; and 6) demographic information. This questionnaire can serve as an instrument for similar survey studies investigating customer preferences and choice behaviors for other products of interest.

## 2. Background

Designing and developing customer-desired products is vital for a company's success in competitive markets. To this end, customer preference modeling is one of the most widely used research methods in marketing science [2,3], and engineering design communities to help identify customer-preferred product features and how customers make tradeoffs among multiple design attributes [2,4]. However, due to data scarcity on customers' social relations, the impact of social influence on the customers' consideration-then-choice decision-making process cannot be explicitly assessed. In current practices, researchers often use synthetic social network data or secondary data, such as online product reviews and social media, to study and infer how social factors influence customers' choice behaviors. Those datasets are not ideal due to limited information on customers' demographics, social ties, and preferences in the consideration and choice stages. Therefore, more accurate and comprehensive data that can address these limitations are needed. In particular, the datasets containing customers' social network data and two-stage preference data in the US household vacuum cleaner market were collected at once in a systematically designed survey.

## 3. Data description

The dataset contains several components, including:

1. The survey instrument used to collect the data in .pdf file format (pdf file for the questionnaire): The survey instruments cover the questions in six major categories:
   a. Customers' two-stage (consideration-then-choice) decision-making process.
   b. Ego-centric social networks, including respondents' general social network (GSN) and product-specific network (PSN) [5]. The GSN is a natural social relation network that captures the people with whom respondents communicate about important issues in their daily lives, such as their spouse, parents, and close friends. The PSN refers to the people with whom respondents have discussed vacuum cleaner purchases, such as their coworkers who have endorsed their purchase, and they may or may not be from respondents' GSN).
   c. Factors (product features and external influence such as advertisement) that influence customers' considerations and choices.
   d. Personal viewpoints about the importance of different aspects of vacuum cleaners (such as quality, appearance, and energy efficiency).
   e. Usage context questions.
   f. Demographic questions.
2. The raw survey data in .csv file format (CSV file for survey data): The survey data contains 251 variables with responses from 1002 respondents. In the survey design, all questions are mandatory. Therefore, no missing values exist, except for instances where respondents chose to respond with "prefer not to say" (in some sensitive demographic questions) or "I don't know" (in some questions related to their social networks). Non-applicable responses are coded as "NULL" and blank values.

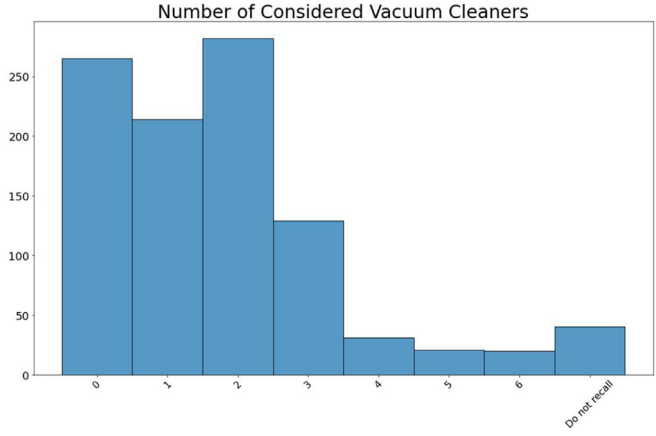


**Fig. 1.** Histogram of the number of vacuum cleaners (other than the purchase one) considered by respondents.

3. The codebook for the raw survey data in .xlsx file format (xlsx file for survey data): The codebook explains the 251 variables included in the survey data file. The codebook lists how each survey question and response option is numerically coded in the raw data and can be used as a guide for navigating the survey dataset.
4. The product feature list in .csv file format (CSV file for product data): The product feature list contains the features of 624 vacuum cleaner products, and each product has 32 variables/features. Missing values, where no online information is available, are coded as "NA". Meanwhile, illegal values, such as runtime for corded vacuum cleaners or navigation path for non-robotic vacuum cleaners, are coded as blank values.
5. The codebook for the product features list in .xlsx file format (xlsx file for product data): The accompanying codebook provides a detailed description of each feature and its data type, as well as the number of missing values for each product feature in the last column.

### *3.1. Survey data summary*

In addition to the data format, we also provide summary statistics of the survey data to help the audience get an overview of the dataset. We primarily use histograms and distribution plots to depict the data in the six major categories outlined above.

Fig. 1 represents the customers' two-stage decision-making process, displaying a histogram of the number of vacuum cleaners (other than the purchase one) all respondents considered.

Fig. 2 shows a histogram of the number of people in each respondent's social network.

Fig. 3 displays a count plot of the factors influencing customers' consideration and purchase stages based on respondents' ranking in the survey data. The plot shows a weighted sum of the respondents' rankings, assigning higher weights to features that received higher rankings. Therefore, the plot presents the weighted feature importance ranking.

Figs. 4-6 depict histograms of personal viewpoints about vacuum cleaners, usage context questions, and demographic questions in the survey, respectively.

### *3.2. Internal consistency check*

Internal consistency is an important measure in survey studies. It is a metric based on the correlation between different questions intended to measure the same construct. In our survey,

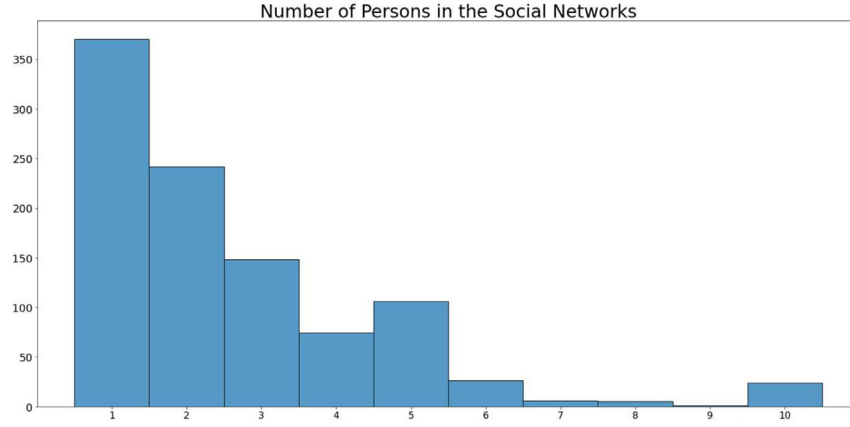


**Fig. 2.** Histogram of the number of people in respondents' social networks.

we aimed to measure customers' personal viewpoints on vacuum cleaners, specifically their demand for high-performing vacuum cleaners. To ensure that the survey instrument was measuring this construct consistently, we assessed the internal consistency of the survey instrument by computing Cronbach's alpha coefficient. Cronbach's alpha is a widely used measure of internal consistency reliability, ranging from 0 to 1, with higher values indicating better reliability. We collected responses from 1002 participants who rated their opinions using 14 Likert-scale items on personal viewpoints about vacuum cleaners. The design of these 14 questions on personal viewpoints was informed by both research interests and expert input. Among these questions, a subset of items, including innovative model, modern technology, reflect lifestyle, environmental-friendly, styling, energy efficiency, after-sale service, high quality, were designed to measure the same underlying construct, namely, the pursuit of better vacuum cleaner quality. The reliability of this subset of items was evaluated using Cronbach's alpha coefficient, which yielded a value of 0.867, indicating excellent internal consistency reliability. The remaining questions in the personal viewpoints section of the survey assessed other aspects of customers, such as brand loyalty, price sensitivity, and susceptibility to social influence, providing additional insights into customer preferences and behavior.

## 4. Experimental design, materials and methods

An overview of the data collection process is shown in Fig. 7. It consists of four major steps, each described in detail below.

**Step 1: Product Database Establishment**

To start the process, we collected information on household vacuum cleaners using two web crawling techniques – Beautiful Soup and selenium in Python. Five primary categories of vacuum cleaner data, i.e., upright, canister, stick, handheld, and robotic vacuum cleaners, were obtained from mainstream online shopping platforms in the US market, including Amazon, Wayfair, Best Buy, Home Depot, and Walmart. After web scraping, data was cleaned to merge data from different sources; meanwhile, the duplicated data and noises were removed too. In the end, 1170 vacuum cleaner products were collected. The collected information includes product title, product image, product model name, SKU (stock-keeping unit), product description, customer rating, customer reviews, and 26 product features (list price, product dimension, weight, manufacture, brand, color, capacity, etc.).

In addition, we extracted product features from online customer reviews to determine the most important (most frequently mentioned) features that shall be included in the survey ques-

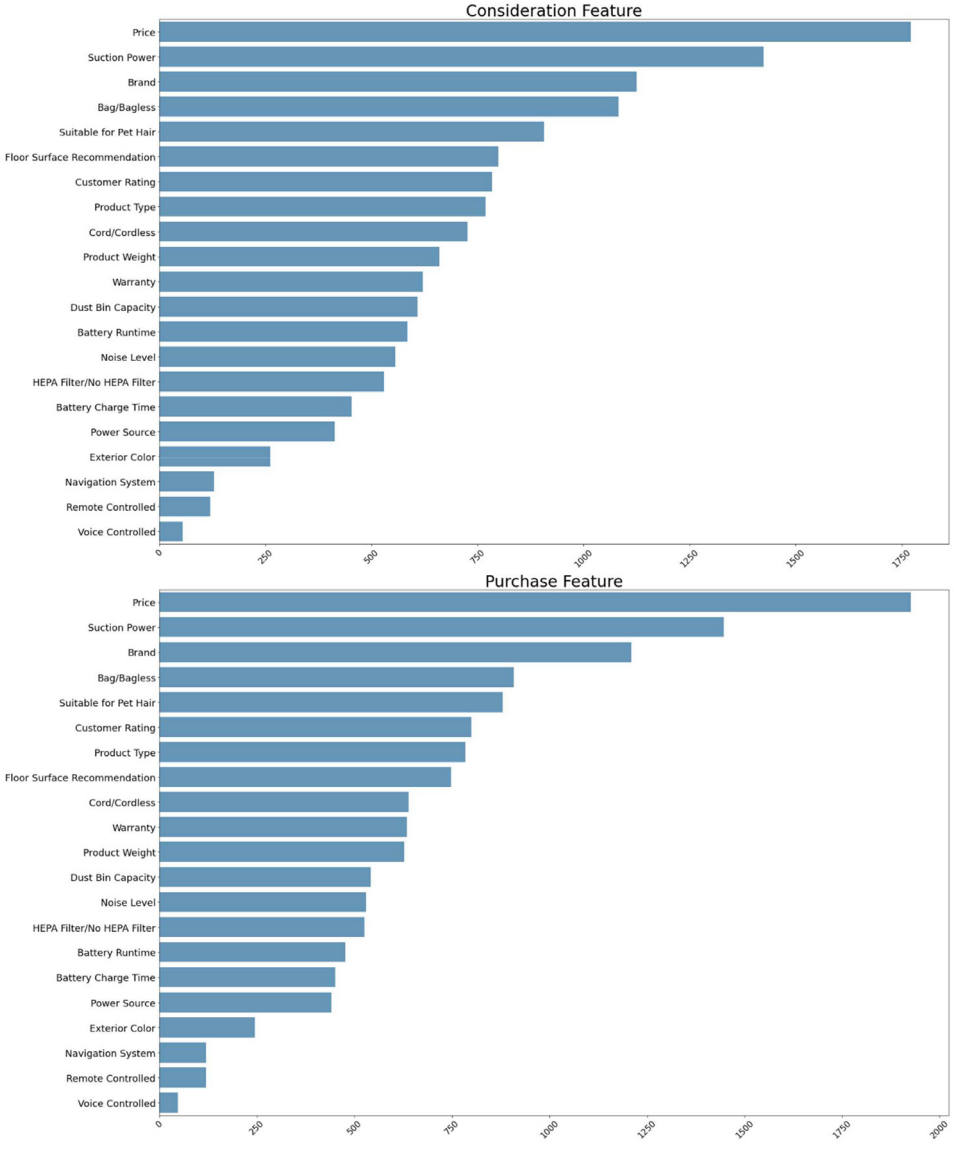


**Fig. 3.** Weighted feature importance rankings reported by respondents for consideration and purchase (choice) stages.

tions. We scraped 60,000 reviews from Amazon (200 reviews for each product) and used a rule-based semi-supervised learning model [6] for extracting features and sentiment/opinions associated with those features. For example, some feature-opinion pairs extracted from the reviews include “strong suction,” “heavy weight”, “annoying cord,” and “loud noise.” After obtaining candidate features from the opinion mining, unrelated features were pruned. The remaining features were then ranked based on their frequency in customer reviews [7]. In the end, we identified 22 important product features based on the results from opinion mining, including attributes such as price, product type, floor surface recommendation, suitable for pet hair, suction power, noise, power source, bag or bagless, cord or cordless, battery charge time, HEPA filter, warranty, brand,

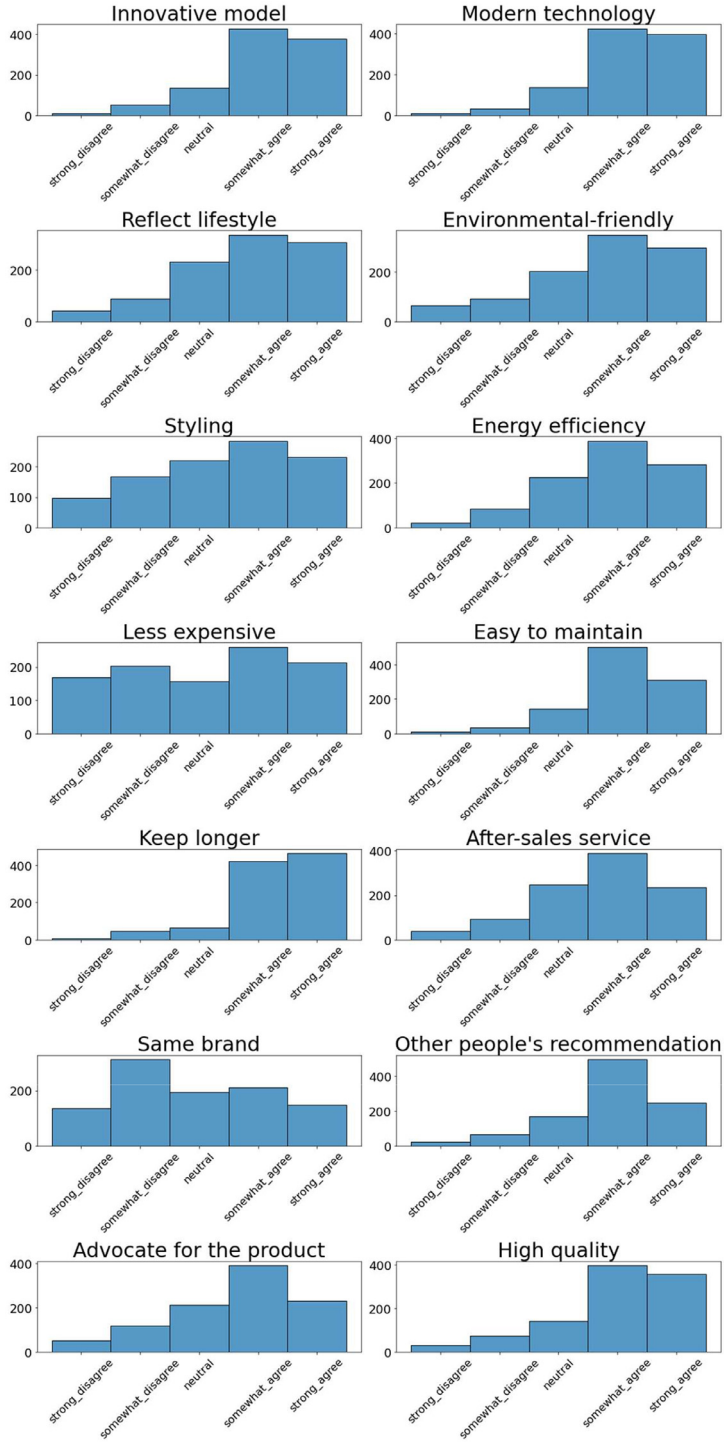


**Fig. 4.** Histogram of variables related to respondents' personal viewpoints about vacuum cleaners.

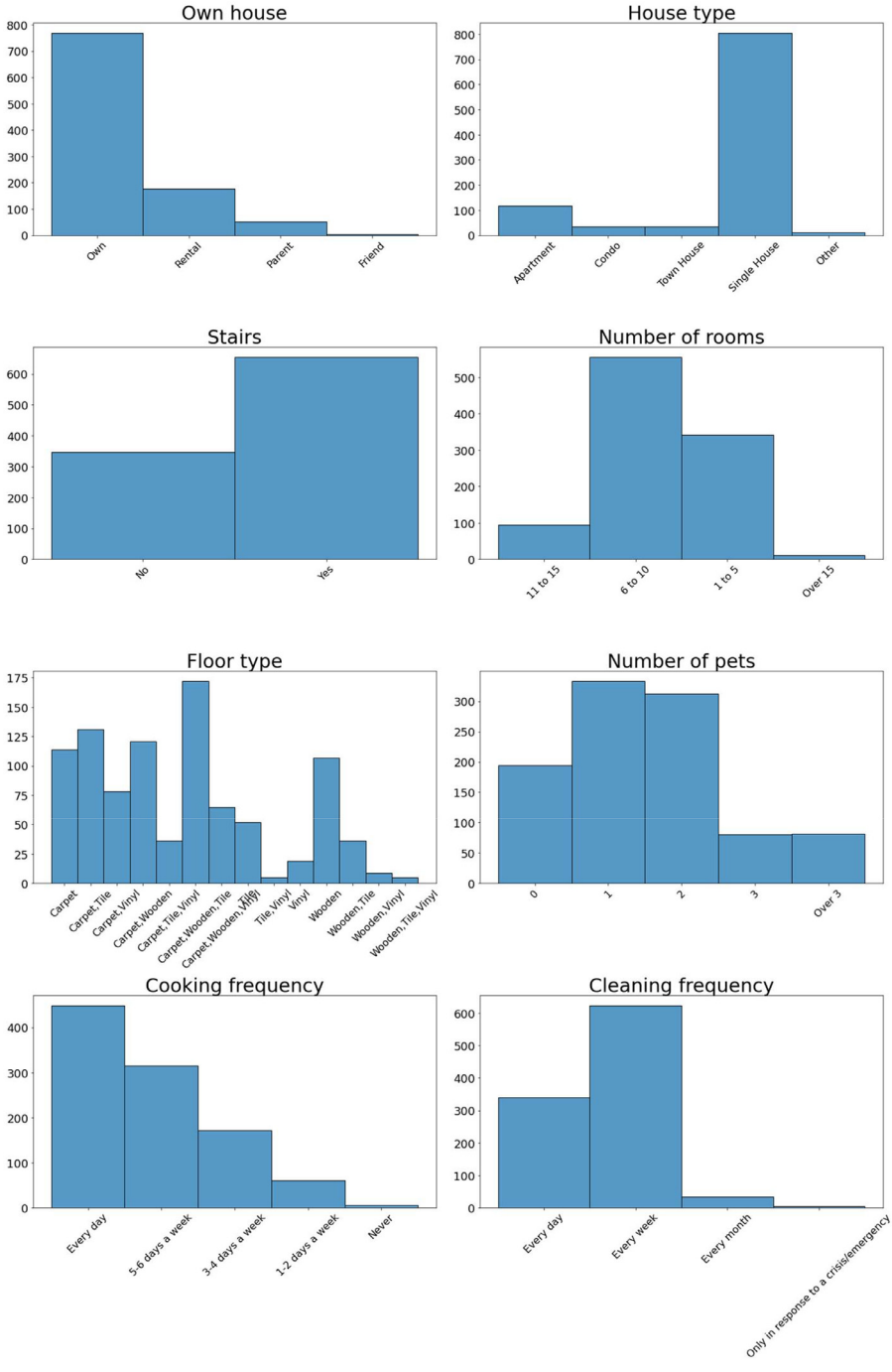


**Fig. 5.** Histogram of variables related to respondents' usage context questions.

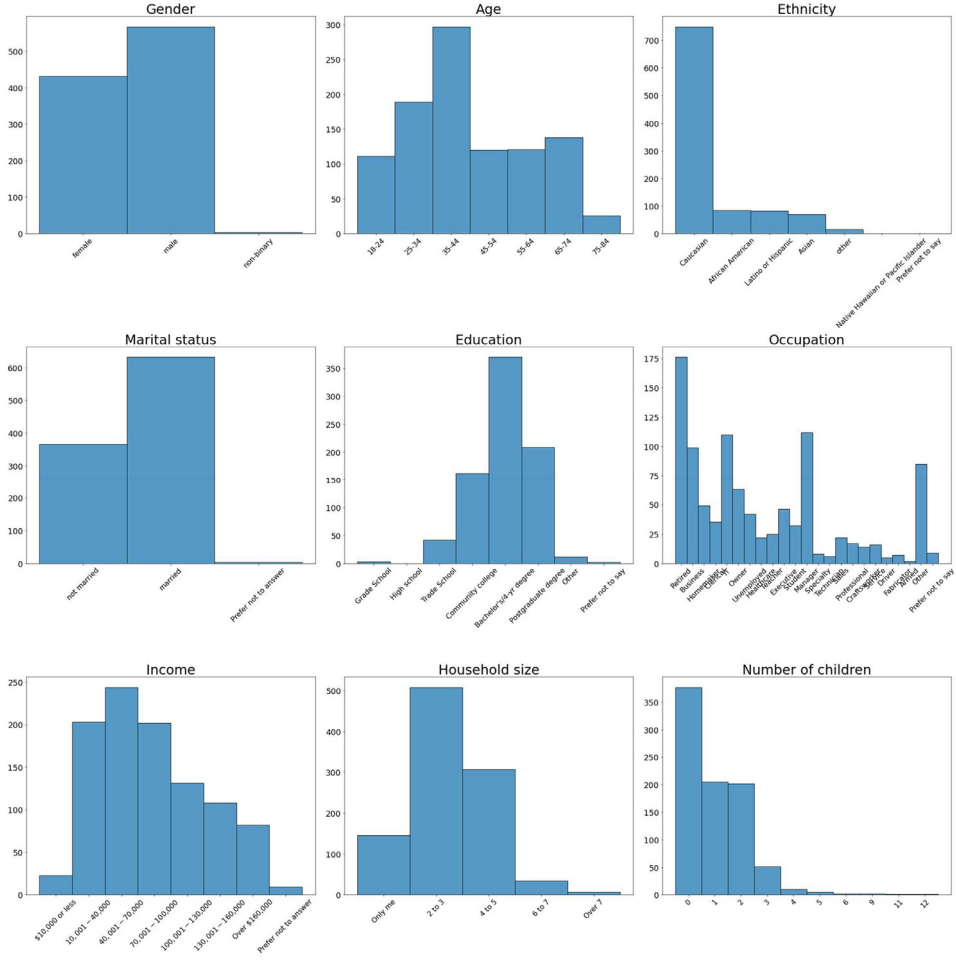


**Fig. 6.** Histogram of variables related to respondents' demographic questions.

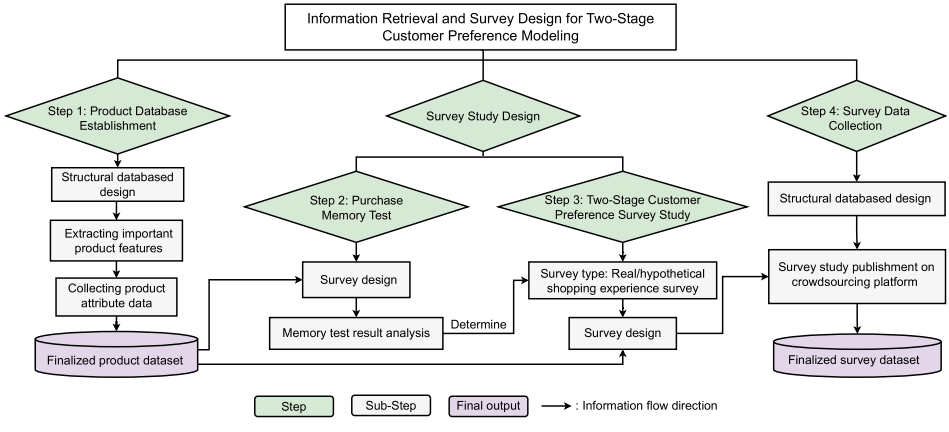


**Fig. 7.** An overview of the data collection process [1].

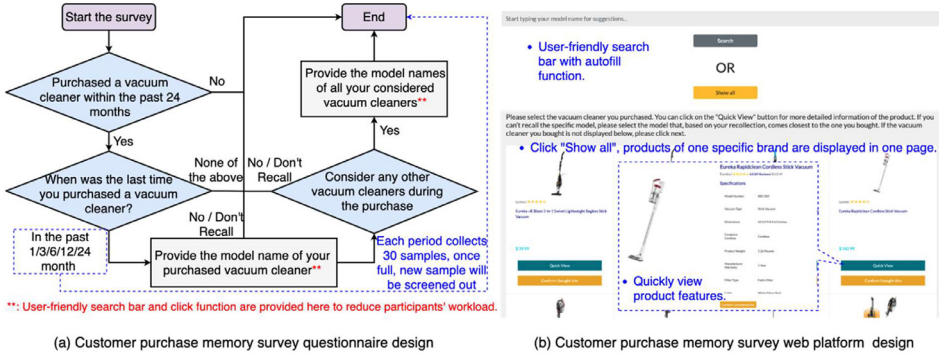


**Fig. 8.** Survey questionnaire flowchart and web platform design for customer purchase memory test [1].

**Table 1**
The sample size of the purchase memory test [1].

| | In the past one month | In the past three months | In the past six months | In the past 12 months | In the past 24 months |
|---|---|---|---|---|---|
| # of people who have purchased a vacuum cleaner | 32 | 34* | 32 | 35 | 8 |

* This number has excluded the number of people who have purchased a vacuum cleaner in the past one month. A similar operation was applied to the other three periods (in the past 6/12/24 months).

color, weight, dimensions, power, capacity, overall customer ratings, and three robotic vacuum cleaner specific attributes (navigation system, voice control, and remote controls).

**Step 2: Customer Purchase Memory Test**

To ensure the credibility of the two-stage customer preference survey study, a memory test was conducted to evaluate customers' abilities to recall their decision-making process while purchasing vacuum cleaners in five different periods: one month, three months, six months, twelve months, and 24 months. This helped us determine the type of survey (i.e., real or hypothetical shopping experience survey) and the appropriate threshold for soliciting participants. In the real one, the survey will be conducted only among the participants who actually purchased the product. In the hypothetical one, participants will be required to complete a survey based on a virtual online shopping experience.

As illustrated in Fig. 8(b), an online survey web was designed and developed. The survey web connected with the product database generated in Step 1 to create a simulated online shopping system. Additionally, we designed user-friendly interfaces, such as the product search bar and product preview, to facilitate participants in identifying the vacuum cleaners they considered and purchased. We collected 30 respondent samples for each period and calculated the proportion of participants who could recall the specific models they considered and purchased. If the proportion exceeded 50%, we considered the customers' memory within that time period to be reliable.

The pilot survey study was conducted on the Cint platform from December 18 to December 21, 2020. Table 1 summarizes the actual collected sample size for the test. It was noted that there were significantly fewer samples for the 24-month scenario than for the other periods, so this scenario was excluded from the proportion calculation. Fig. 9 indicates that 62% of customers who purchased a vacuum cleaner in the past three months can recall their purchases and considerations, meeting the 50% threshold. However, focusing solely on customers who purchased vacuum cleaners within the past three months may not yield enough samples for the subsequent two-stage customer preference survey in Step 3. To strike a balance, the survey study in Step 3 was extended to include customers who made purchases within the past six months as they had a high recall ratio for purchase (75%) and an acceptable ratio for a recall of

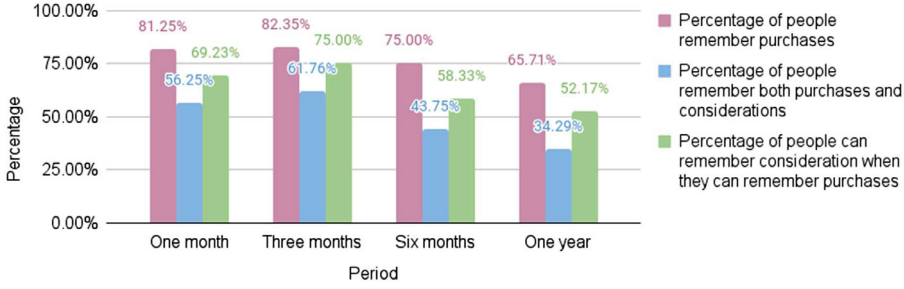


**Fig. 9.** The ratio of participants who can recall the purchased or considered vacuum cleaners [1].

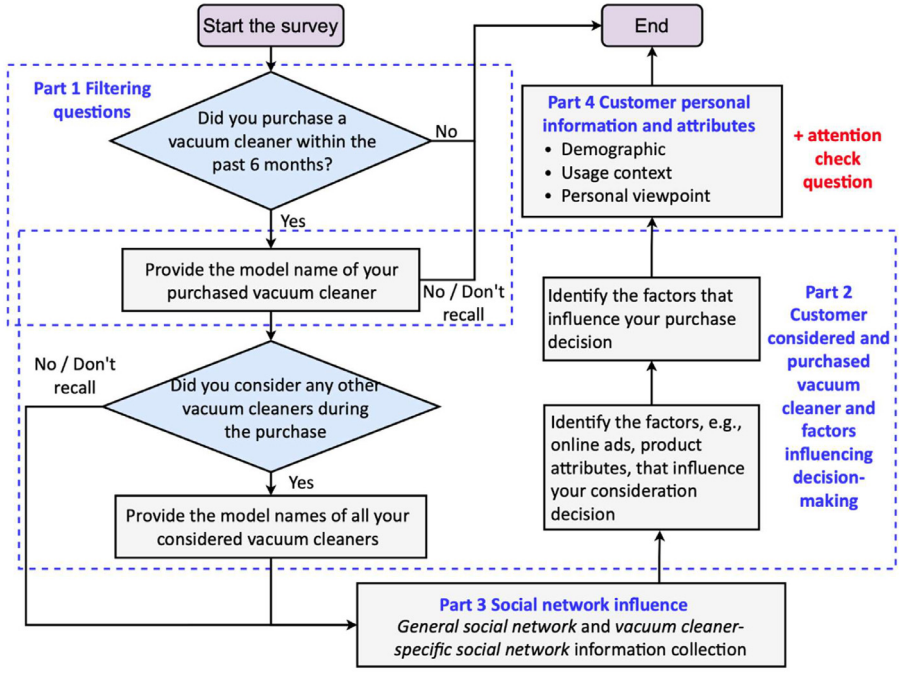


**Fig. 10.** Two-stage customer preference survey questionnaire flowchart.

both purchase and consideration (43.75%). Therefore, according to the memory test results, we decided to conduct a study on customers' revealed preferences and recruited participants who had purchased a vacuum cleaner in the past six months.

**Step 3: Two-Stage Customer Preference Survey Questionnaire Design**

Step 3 involves designing the two-stage customer preference survey questionnaire. As shown in Fig. 10, the questionnaire consists of four major parts. **Part One** includes two filtering questions to collect respondents' vacuum cleaner purchase decisions, which are the most important information we want to collect. Only the respondents who purchased a vacuum cleaner within the past six months and could recall the products they purchased were allowed to participate in the rest survey.

In **Part Two**, the online survey web shown in Fig. 8(b) was used to collect participants' historical consideration and choice data. They were asked to provide information about the type, brand, and exact models of vacuum cleaners they have considered and purchased, as well as the

**Table 2**
The total number of participants and the number of complete responses received in each phase. Participants' responses could be removed due to: 1) early screening: Participants who did not purchase a vacuum cleaner, disagreed with the survey agreement, or did not specify their purchased vacuum cleaners, were screened early in the process; 2) incomplete survey: Participants who did not complete the survey in its entirety were excluded; 3) attention check failures: participants who did not pass the attention check questions were excluded; 4) suspicious cheating: Instances of suspicious behavior, such as inputting irrelevant words or sentences in text boxes and consistently providing the same answer (e.g., "Strong Agree") to all personal viewpoint questions, led to participant removal.

| | Phase 1 | Phase 2 | Phase 3 | Phase 4 |
|---|---|---|---|---|
| # of participants | 828 | 1263 | 2002 | 2492 |
| # of complete responses | 101 | 220 | 292 | 410 |

top-rated design attributes (product features) that influenced their choice-making. Participants could rank these attributes by dragging them from a list of features identified by the feature selection algorithm introduced in Step 1 to the corresponding text boxes.

In **Part Three**, we design questions to collect participants' social network data. This was relevant because social networks can influence consumers' purchase decisions. Participants were asked to provide information on their general social networks (GSN) as well as product-specific social networks (PSN), both of which have the potential to influence participants' choice behaviors. Each participant was asked to provide information for at least one and up to five individuals in their GSN with whom they discuss daily matters. Additionally, they were asked to provide information for up to five other individuals in their PSN with whom they had discussed the vacuum cleaner purchase. Therefore, each participant can nominate up to a total of ten different people in their social network for the study. These individuals' demographic data and their contact frequencies with the respondents were also recorded.

**Part Four** aimed to collect personal information and general preferences of the participants, such as their demographics and viewpoints about vacuum cleaners. Additionally, this part of the survey focuses on understanding the product usage context of the participants, including how often they use the vacuum cleaner and where they use it. To ensure the quality of the survey data, we employed several strategies [8]:

1) Designed and implemented attention check questions;
2) Organized questions by placing important ones first and less important ones last;
3) Made questions mandatory to avoid missing data, i.e., participants could not proceed to the next stage unless answering all the required questions on the current page.
4) Conducted both internal and external pilot studies to collect feedback on the questionnaire;
5) Incorporated experts' inputs and feedback from multiple disciplines, including engineering design, social science, and psychological science.

**Step 4: Survey Data Collection**

We launched our survey on Cint, a digital insights gathering platform with quality assurance mechanisms such as artificial intelligence (AI)-driven fraud detection system. To ensure reliable data storage, the survey data was automatically saved in an SQL database on pgAdmin, with a structured column sequence. This database had been configured to communicate effectively with the survey website. To acquire more results, the survey was distributed to different groups, such as those who had recently purchased a vacuum cleaner or those who were interested in home decoration and home appliances. Meanwhile, to mirror the real market, a quota sampling technique [9] was used to match the age distribution of the US census. The survey was conducted over two months, from April 25 to June 25, 2021, with the aim of collecting approximately 1,000 complete responses. To improve the reliability of data collection, we divided this data collection process into four phases. Each phase targeted an equidistant increase, with goals set at 100, 200, 300, and 400 complete responses from Phase 1 to Phase 4. Table 2 provides a summary of the actual number of participants and the complete responses obtained in each phase. After obtaining a total of 1023 complete responses, a subsequent manual check identified 21 responses related to hard-to-find vacuum cleaners, prompting their removal. Finally, a total of 6585 partic-

ipants attended the survey and 1002 complete responses were received, with a completion rate of 15.21%.

From the data collected, we identified 624 unique vacuum cleaner models that had been either considered or purchased by the respondents. However, given that the scrapped product attribute data in Step 1 included a considerable number of missing values, we conducted an additional round of manual data collection to address the missing value issue. This manual collection involved gathering information from various sources, such as product specifications and manuals, the brand's official online stores, and expert performance review reports available online.

## Limitations

The potential limitations of this dataset are twofold. First, it comprises survey data collected during a particular time period, rendering it unsuitable for temporal analysis. Consequently, it does not support the study of evolutionary changes in customer preferences. Second, the social network data, including demographic information from individuals and details about their purchased vacuum cleaners, are based on self-reporting by respondents. This aspect introduces the possibility of data noise and inaccuracies.

## Ethics statement

Ethical approval of this study has been granted by the University of Arkansas with protocol number 2004263007. All participants were given informed consent and had the freedom to withdraw from the survey at any time. We prioritize the privacy rights of participants, and as such, the data collected does not reveal individual identities through their responses. To maintain complete anonymity, no information was included that could be used to identify any participant.

## Data Availability

Survey Data on Customer Two-Stage Decision-Making Process in Household Vacuum Cleaner Market (Original data) (Texas Data Repository).

## CRediT Author Statement

**Yinshuang Xiao:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization; **Yaxin Cui:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization; **Nikita Raut:** Methodology, Software; **Jonathan Januar:** Writing – review & editing, Methodology; **Johan Koskinen:** Writing – review & editing, Supervision; **Noshir Contractor:** Conceptualization, Methodology, Writing – review & editing, Supervision; **Wei Chen:** Conceptualization, Methodology, Writing – review & editing, Supervision; **Zhenghui Sha:** Conceptualization, Methodology, Writing – review & editing, Supervision.

## Acknowledgements

We acknowledge Lada Nuzhna, Olga Lew-Kiedrowska, Laith Kassisieh, Neelam Modi for their assistance in data management on Cint and/or the inputs during research meetings. We also greatly acknowledge the funding support from NSF CMMI #2005661 and #2203080.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] Y. Xiao, et al., Information retrieval and survey design for two-stage customer preference modeling, Proc. Design Soc. 2 (May 2022) 811–820, doi:10.1017/pds.2022.83.
[2] C. Pescher, M. Spann, Relevance of actors in bridging positions for product-related information diffusion, J. Bus. Res. 67 (8) (Aug. 2014) 1630–1637, doi:10.1016/j.jbusres.2013.09.005.
[3] A. Stankevich, Explaining the consumer decision-making process: critical literature review, J. Int. Bus. Res. Mark. 2 (6) (2017) 7–14, doi:10.18775/jibrm.1849-8558.2015.26.3001.
[4] Z. Sha, V. Saeger, M. Wang, Y. Fu, W. Chen, Analyzing customer preference to product optional features in supporting product configuration, SAE Int. J. Mater. Manuf. 10 (3) (2017) 2017-01–0243, doi:10.4271/2017-01-0243.
[5] K.E. Campbell, B.A. Lee, Name generators in surveys of personal networks, Soc. Networks. 13 (3) (Sep. 1991) 203–221, doi:10.1016/0378-8733(91)90006-F.
[6] T.A. Rana, Y.N. Cheah, A two-fold rule-based model for aspect extraction, Expert. Syst. Appl. 89 (Dec. 2017) 273–285, doi:10.1016/J.ESWA.2017.07.047.
[7] R. Rai, Identifying key product attributes and their importance levels from online customer reviews, in: Volume 3: 38th Design Automation Conference, Parts A and B, American Society of Mechanical Engineers, Aug. 2012, pp. 533–540, doi:10.1115/DETC2012-70493.
[8] R. Flowerdew and D. M. Martin, Eds., *Methods in Human Geography*. Routledge, 2013. 10.4324/9781315837277.
[9] S. Sudman, Probability sampling with quotas, J. Am. Stat. Assoc. 61 (315) (Sep. 1966) 749–771, doi:10.1080/01621459.1966.10480903.