Eliciting Preferences to find your Perfect Laptop: A Usability Study

Jhoan Almeida[0000−0002−6225−9988] and Ixent Galpin[0000−0001−7020−6328]

Facultad de Ciencias Naturales e Ingeniería
Universidad de Bogotá - Jorge Tadeo Lozano
Bogotá, Colombia
{jhoans.almeidac,ixent}@utadeo.edu.co

Abstract. Most e-commerce websites that sell technological items provide a simplified Query-by-Example preference elicitation approach to enable users to search for the product that they desire by specifying constraints over criteria and filtering results. Such an interface may be adequate when the technical specification is known. In this work, the usability of an alternative preference elicitation approach, Pairwise Comparisons, that enable users to specify trade-offs between various potentially conflicting criteria, is compared and contrasted to the traditional Query-by-Example approach.
This is done by implementing two search tools that implement each preference elicitation approach to collect data about user behaviour. An experiment is carried out to evaluate performance and usability to search for laptops over a web scraped data set for varied tasks. It is found that while Query-by-Example tends to be preferred for cases where the technical specification criteria are precisely specified, the result sets generated for the Pairwise Comparisons approach are preferred when it is less clear what the criteria values should be. As such, the Pairwise Comparisons preference elicitation is deemed to be a worthwhile complementary approach for such e-commerce websites.

Keywords: Usability · Preference Elicitation Approaches · Pairwise Comparisons · Query-by-Example

1 Introduction

Choosing which laptop to buy can be a daunting prospect. For many people, the acronyms and technical specifications may be hard to parse. Even experienced computer users are likely to find that technology has moved on since they purchased their last device, and that getting up to date with the latest developments is a cumbersome chore. To aid with this difficulty, there are numerous sites on the Web (e.g., http://www.which.co.uk or http://www.techradar.com) that purport to advise consumers about the best computers to buy, and in order to contend with the high number of products available may provide a ranking of models for different prices. This is often coupled with tips to take into consideration when making a purchase that has significant implications for most household budgets.

Most e-commerce sites such as Amazon or Ebuyer incorporate search tools that assume that the purchaser knows the technical specifications of the laptop that they wish to buy. Typically, they are presented with with an interface that enables them to filter in terms of attributes that denote laptop characteristics. For example, a user may specify requirements in terms of the minimum RAM available, or screen size required, usually dictated by the anticipation of tasks that they wish to

perform. As such, a user may filter for laptops whose RAM is greater than 16GB if the laptop is required for certain memory intensive tasks. This approach, which essentially comprises a subset of the *Query-by-Example* (QBE) graphical query language proposed by [13], has a demonstrated track record of being an intuitive way for users to find the information that they require [12].

However, using a filtering approach such as QBE may be less effective if a user is unsure about the values that are required for certain criteria, e.g., RAM. Furthermore, trade-offs may present themselves among the various criteria associated with a laptop. For example, the lighter a laptop is, the more expensive it is likely to be. Such nuances cannot be captured using a preference elicitation approach such as QBE. Other elicitation approaches such as Pairwise Comparisons (PCs) have been proposed [7] that enable the relative importance of pairs of criteria to be specified by a user, rather than constraints in terms of specific thresholds.

In the paper, we carry out an exploration of the effectiveness of QBE and PCs as elicitation techniques and measure factors including performance and usability. We consider two categories of tasks: those in which the criteria threshold values to consider are *explicit*, and thus arguably lend themselves better to the QBE approach, and those in which the criteria are implicit, and which may therefore lend themselves better to PC elicitation. The contributions of this paper are:

1. The implementation of two search tools, one for QBE and the other one for PC, to perform data collection about performance and usability;
2. The design of an experiment to measure the performance and usability of both preference elicitation techniques, using the search tools created in (1);
3. The analysis of the data collected from the experiment in (2).

This paper is structured as follows. Section 2 presents related work. In Section 3, we present a description of the tools evaluated, and the associated paradigms employed by each tool. The usability experiment is described in Section 4. Subsequently, in Section 5 we describe results. We present a discussion in Section 6. Section 7 concludes.

2 Related Work

The verbal Saaty scale, shown in Table 1, is a widely used scale used to enable users to denote strength of preference between criteria in terms of pairwise comparisons (PCs) [7]. Although this preference elicitation approach was originally conceived for the Analytic Hierarchy Process (AHP) methodology [8], it has been used as a preference elicitation more widely in diverse computing applications. A number of usability studies have been carried out to show its effectiveness in relation to other techniques. For example, Galpin *et al.* [3] and Abel *et al.* [1] show that it is able to obtain improved performance and usability in the domain of source selection (the problem in data integration whereby a subset of data sources needs to be chosen) when compared to constrained optimisation as an elicitation technique. Leon-Medina *et al.* [5] also demonstrate that they are preferred by users in the domain of data cleaning.

There has also been discussion about the effectiveness of pairwise comparisons as a preference elicitation approach in the field of recommender systems. Kalloori *et al.* [4] consider how pairwise preferences may be elicited in a Recommender System and effectively combined with absolute ratings, and demonstrate that pairwise preferences enable better recommendations to be made to users. They conclude that the most effective preference elicitation approach is inextricably linked to the user's personality. Sepliarskaia *et al.* [11] also uses relative preference questions combined with a latent factor model to improve recommendation quality.

The usability of QBE has been explored since its inception. Thomas *et al.* [12] compare three approaches: the "English-like" IQF and SQL languages, and the graphical QBE language. They find that QBE is effective as users tend to not confuse disjunctive and conjunctive queries, and easily learned by novice users. It is likely for this reason that subsets of QBE are used by search tools on e-commerce websites pervasively.

While usability studies have been carried out about QBE and PC preference elicitation techniques in isolation, the authors are not aware of any usability studies in which these two techniques are directly compared to one another.

Verbal Preference Strength	**Numerical**
Equal importance	1
Weak or slight importance	2
Moderate Importance	3
Moderate plus	4
Strong Importance	5
Strong plus	6
Very strong importance	7
Very very strong	8
Extreme importance	9

Table 1. The verbal Saaty scale [7]

3 Search Tools

In this section we describe the search tools that we implemented using QBE and PC preference elicitation approaches. The search tools were implemented using Python 3.8 with the PyCharm IDE and Django[1] framework. Real data concerning 90 laptops was scraped from the Mercadolibre website[2] using the scrapy[3] Python library. Mercadolibre is an e-commerce website similar to e-bay and is currently the company with the highest market capitalisation in Latin America [6]. Data was scraped from "official" vendors because it was found to be of higher quality. Not only was it significantly more complete, it was also available in a structured format rather than the free-text format from non-official vendors. Data cleaning was carried out using the pandas[4] y NumPy[5] Python libraries. The data set was stored in a PostgreSQL database and made available for querying by the search tools using the Python pg2psyco library[6].

[1] https://www.djangoproject.com/
[2] https://www.mercadolibre.com.co/
[3] https://scrapy.org/
[4] https://pandas.pydata.org/
[5] https://numpy.org/
[6] https://pypi.org/project/psycopg2/

3.1 Query-by-Example Tool

Figure 1 presents an example of the user interface for the QBE tool. The preference elicitation interface allows a user to enter one or more constraints or filter conditions to specify the technical specification required. The constraints must be given in terms of a criterion, a comparison operator, and a threshold value for the criterion. The criteria available are {brand, price, RAM, screen size, screen type, storage, webcam, weight}. For numerical criteria, there are five different comparison operators to choose from $\{<, \leq, =, \geq, >\}$ however these are presented verbally (e.g, "greater than or equal to") in the drop-down menu to ease of understanding of non-mathematically minded people and for consistency with the Pairwise Comparisons tool (described ahead). The threshold value depends on the criterion selected, e.g., for *Brand* the options available are {Acer, Asus, Compumax, Dell, HP, Huawei, Lanix, Lenovo, Toshiba}. The Add and Remove buttons allow a user to add as many constraints as desired.

When the Evaluate button is clicked, a result set of up to five laptops that meet the constraints given are shown. These are chosen at random from the database. It is possible that less than five results are shown or indeed, none, if the constraints are too restrictive. For the purposes of the experimental evaluation, the user is only given one opportunity to carry out the search, and then is asked to state whether he or she likes the result or not.

In Figure 1(a), the user seeks a laptop with at least 4GB RAM and large display size. However, the user has a maximum budget of 6 million Colombian pesos. All the results in Figure 1(b) meet these constraints.

3.2 Pairwise Comparisons Tool

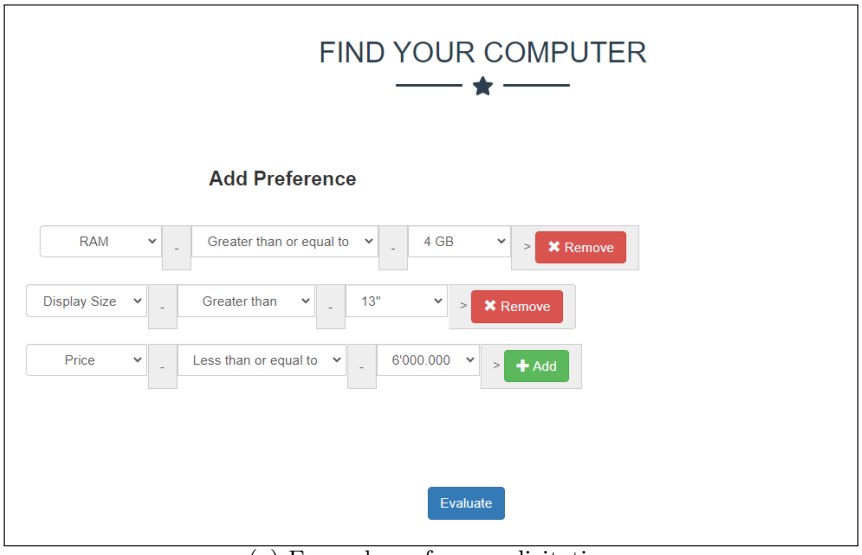
Figure 2 presents an example of the user interface for the PC tool. The interface aims to appear visually as similar as possible to the QBE tool in order to ensure fairness of the comparison we carry out in the subsequent experimental evaluation. This is achieved by, as far as possible, using the same type of user interface elements and layout in both cases.

The preference elicitation interface allows a user to enter a series of preferences in terms of two criteria and an operator to express the degree of preference between the left-hand and right-hand side criteria. The set of criteria used are the same as for the QBE tool. However, the preference operator is a subset of the verbal Saaty scale shown in Table 1. Five points in the scale are chosen, so that the number of scale points is the same as with the QBE tool. Unlike the QBE tool, no threshold values are given, as the idea is to specify the relative importance of criteria.

Once the user preferences are elicited, the well-known Analytical Hierarchy Process (AHP) methodology [8] is applied in order to select the laptops which best meet the user preferences. In Figure 2(a), the user gives a series of preferences, and the result set produced is shown in Figure 2(b).

4 Experiment Design

We designed an experiment to measure the performance and usability of both preference elicitation techniques. For this, the tasks shown in the Table 2 were defined. This table shows two types of tasks: those that stipulate an explicit value for the thresholds of the criteria, and those that express the user's requirements without specifying specific values for the criteria. It is expected that for tasks whose criteria values are explicitly defined, the requirements can be directly expressed using



(a) Example preference elicitation

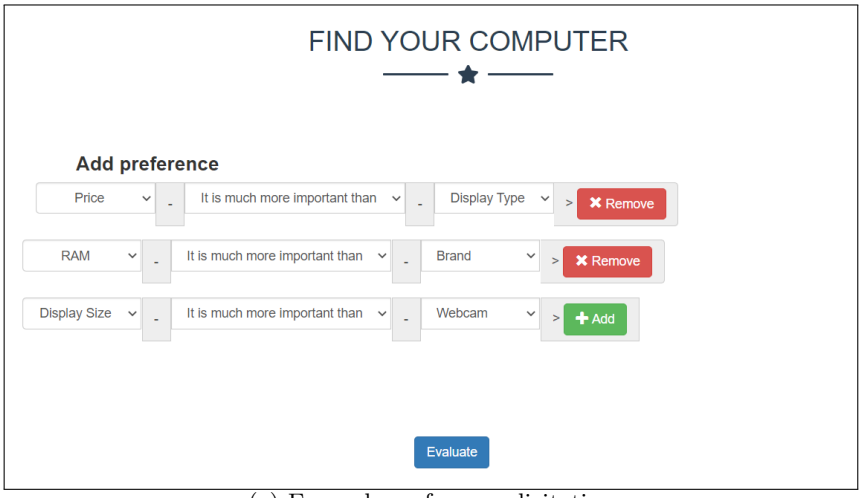
(b) Example results

Fig. 1. The *Query-by-Example* Interface

QBE. In contrast, for task requirements where only specific values are specified, the expectation is that these can be expressed using PC more easily.

The structure of the experiment is defined as follows:

- A short two-minute tutorial is carried out explaining each of the elicitation techniques and the steps to be followed;
- The participant carries out the four tasks in Table 2 using Technique 1. For each task, the user gives the requirements according to their interpretation, and the tool returns the result sets. The user must then indicate whether he is satisfied with the results, indicating whether they are deemed to be useful or not.
- The participant answers four questions about the usability of Technique 1;



(a) Example preference elicitation

(b) Example results

Fig. 2. Pairwise comparisons

- The participant carries out the four tasks in Table 2 using Technique 2;
- The participant answers four questions about the usability of Technique 2;
- The participant answers a question comparing the usability of both elicitation techniques, giving a qualitative answer.

Two groups of participants of equal size are created. For the first group, Technique 1 corresponds to Query-by-Example, and Technique 2 to Pairwise Comparisons. For the second group, the reverse is done. This is to avoid giving an advantage to one of the techniques over the other, taking into account that the second technique will always have an advantage since participants become more familiar with the interface. Two versions of a guide with instructions were created for each group, with the respective links to the tools and the usability questions.

The metrics used for evaluation aim to consider two aspects, user performance and usability. *User Performance* is related to the F-measure of the result set obtained for a given preference elicitation.

Task	Description	Criteria values
1	I am a video game designer and I need to select a computer for my work. The price and weight are not important to me.	Implicit
2	I manage a team for a large company. Usually I need a computer to use MS Office applications, to send emails, web-browsing and video-conferencing. It is important that it not be too heavy as I travel frequently, and that the price is between the 3 to 5 million Colombian pesos assigned in the budget.	Explicit
3	I am a person without much technical knowledge and I need a computer to get in touch with my children, view photos and use social networks. I don't want to spend much money.	Implicit
4	I am a web page developer and I need a computer that supports several programs in parallel to give good customer support, possibly with 4GB of RAM, or higher, and a screen size larger than 13 inches. My maximum budget is 6 million Colombian pesos.	Explicit

Table 2. Tasks given to participants of the usability experiment

F-measure is a widely used measure within information retrieval that provides a measure of accuracy combining both precision and recall [10]. It also enables result sets that may be of different sizes to be compared in a succinct manner. The F-measure obtained for a given result set $\mathbf{R}$ for a task T and elicitation e, denoted $F(\mathbf{R}_e^T)$, is calculated thus:

$$F(\mathbf{R}_e^T) = 2 \cdot \frac{\mathsf{Precision}(\mathbf{R}_e^T) \cdot \mathsf{Recall}(\mathbf{R}_e^T)}{\mathsf{Precision}(\mathbf{R}_e^T) + \mathsf{Recall}(\mathbf{R}_e^T)} \tag{1}$$

where $\mathbf{R}_e^T$ is result set obtained using elicitation e and

$$\mathsf{Precision}(\mathbf{R}_e^T) = \frac{|TP_e^T|}{|TP_e^T \cup FP_e^T|} \tag{2}$$

and

$$\mathsf{Recall}(\mathbf{R}_e^T) = \frac{|TP_e^T|}{|TP_e^T \cup FN_e^T|} \quad . \tag{3}$$

The set of true positives TP_f^T is defined thus:

$$TP_e^T = \{r | r \in \mathbf{R}_e \wedge r \in \mathbf{F}_T\} \tag{4}$$

where $\mathbf{F}_T$ is the set of true positives, i.e., the rows in the results which are deemed to be adequate according to the natural language description for task T. By true positive is meant a record corresponding to a laptop computer which deemed to be suitable for the task described. For example, for Task 2 in Table 2, which specifies a budget of between 3-5 million, an example of a true positive would be a laptop computer in that range. Conversely, a laptop with the price outside that range would be deemed to be a false positive.

For usability, we consider the two metrics to measure subjective user preference. Firstly, we compute the frequency of *likes* vs. the frequency of *dislikes* of the result sets obtained for each

Question	Type
I found the system unnecessarily complex.	Negative
I would imagine that most people would learn to use this system very quickly.	Positive
I thought the system was easy to use.	Positive
I needed to learn a lot of things before I could get going with this system.	Negative

Table 3. Subset of SUS questions employed

query. Furthermore, we calculate a usability score based on the well-established System Usability Scale (SUS) [9]. We use a variant of the SUS score which only involves four questions, in a similar fashion to Leon-Medina *et al.* [5]. The four questions used are shown in Table 3. We collect the responses using the Likert scale [2], the the options strongly disagree, slightly disagree, neither agree nor disagree, slight agree and strongly agree. The usability score is each technique calculated as follows:

- The positive questions take the value assigned by the user minus one.
- The negative questions are 5 minus the value assigned by the user.

The individual scores for the questions are summed. Subsequently, a score in the range of 0 to 100 is computed:

$$\mathrm{U} = \sum_{i=1}^{N} (score_i) \times \frac{100}{4N} \tag{5}$$

where N is the number of questions (four in this case), $score_i$ is the score awarded to the ith question, and the constant 4 represents the maximum score for any given question.

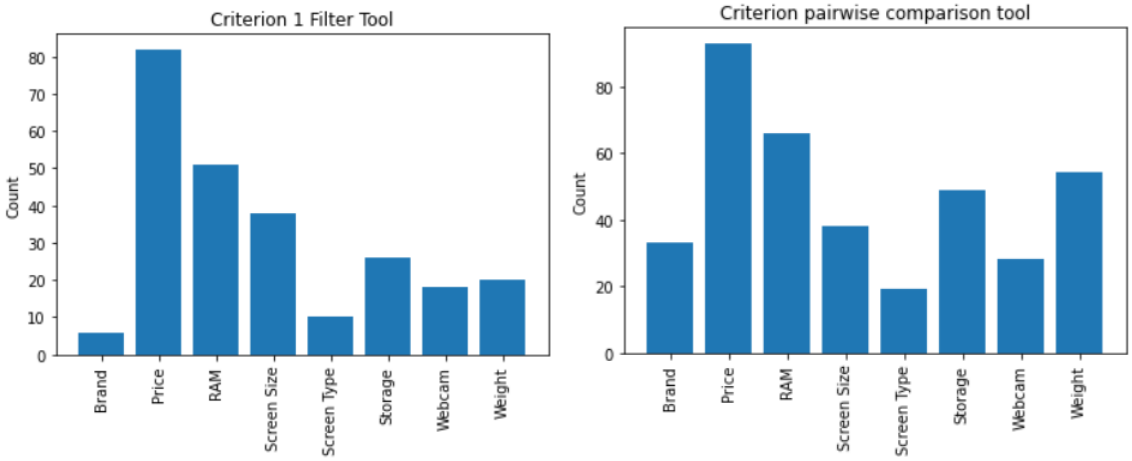
5 Results

Twenty volunteer participants were recruited to carry out the usability experiment described. The majority of participants were Computer Science students at Jorge Tazeo Lozano University in Bogotá, Colombia. The remainder were friends and family members of the authors. The people who took part had a diverse range of IT skills knowledge, ranging from elementary to advanced.

Frequency of Criteria Employed Figure 3 shows the frequency of use for the criteria Query-by-Example and Pairwise Comparisons respectively. The price and RAM criteria are most frequently used for both elicitation approaches. The third most frequently used criterion is *Screen Size* and *Weight* for Query-by-Example and Pairwise Comparisons respectively. The least frequently used criteria are *Brand* for Query-by-Example and *Screen Type* for the Pairwise Comparisons elicitation approach. It is broadly observed that the criteria employed reflect the descriptions of the tasks given in Table 2.

Frequency of Comparison Operators Employed Figure 4(a) shows the frequency of use of the types of comparison operators for the Query-by-Example approach. The *greater than or equal to* or *less than or equal to* comparison operators are the most frequently used, a direct reflection of the tasks in the experiments which tend to mention inclusive ranges when thresholds are given.

In contrast, Figure 4(b) shows that users prefer to provide their preferences in positive terms (i.e., a criterion X being preferred to Y) rather than negative terms (i.e., Y being *less* preferred to X). This observation suggests that the PC interface may be simplified to exclude the negative options, which appear to be less intuitive.



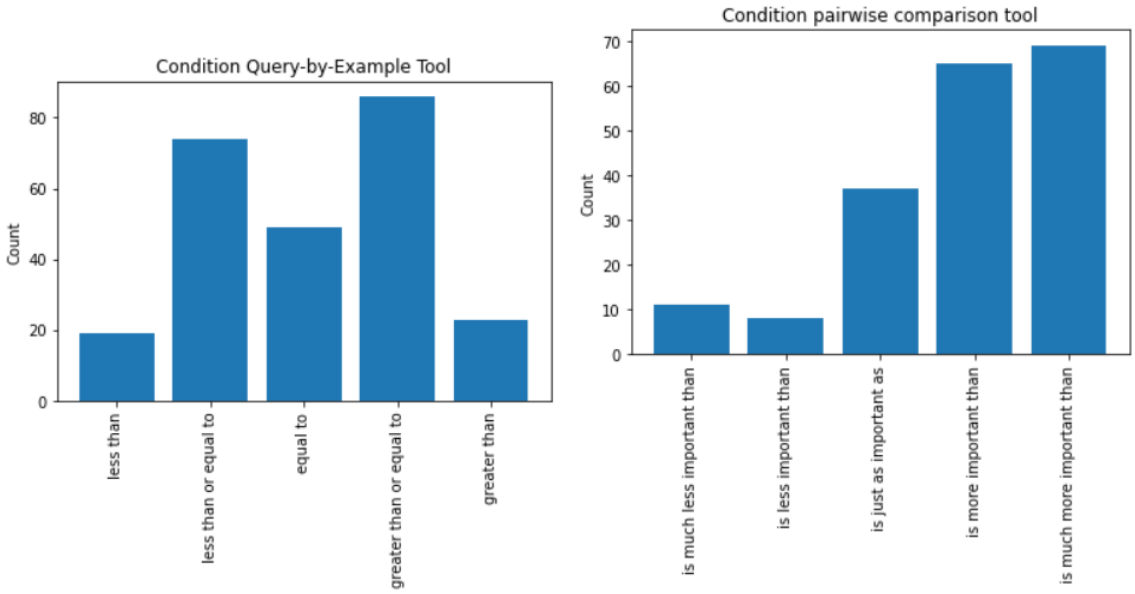
(a) Query-by-Example (b) Pairwise Comparisons

Fig. 3. Frequency of Criteria Use

User Performance Figure 5 presents scatterplots showing precision vs. recall for each of the four tasks. Overall, it is observed that precision and recall tend to be higher for the Query-by-Example elicitation approach rather than Pairwise Comparisons. For Tasks 1 and 3, while it is observed that most users obtain full precision for Query-by-Example, the recall obtained is quite variable. This is most likely due to contraints over the criteria at times leading to less than five results being returned by the tool.

Usability Score Figure 6 shows the Usability Scores obtained for each person. On average, the Usability Scores were 75.3 for Query-by-Example, and 72.5 for Pairwise Comparisons. The difference is negligible and a paired t-test shows that the difference between the two distributions is not significant.

User result set evaluation Figures 7 and 8 show user result set evaluation for the Query-by-Example and Pairwise Comparison elicitations approaches respectively. It is observed that for Tasks 1 and 3, with implicit criteria, that participants are happy with the result set given by the Pairwise Comparisons elicitation approach 46.2% and 13.5% resp. more often than the Query-by-Example approach. A paired two-tailed t-test shows that this is very significant in the case of Task 1 ($p = 0.0140$) and less so in the case of Task 3 ($p = 0.336$). On the other hand, for Tasks 2 and 4, which have explicit criteria, participants are happy with the result set given by Query-by-Example approach 4.05% and 6.86% more often that with the Pairwise Comparisons. However, these results are not statistically significant.



(a) Query-by-Example (b) Pairwise Comparisons

Fig. 4. Frequency of Comparison Operator Use

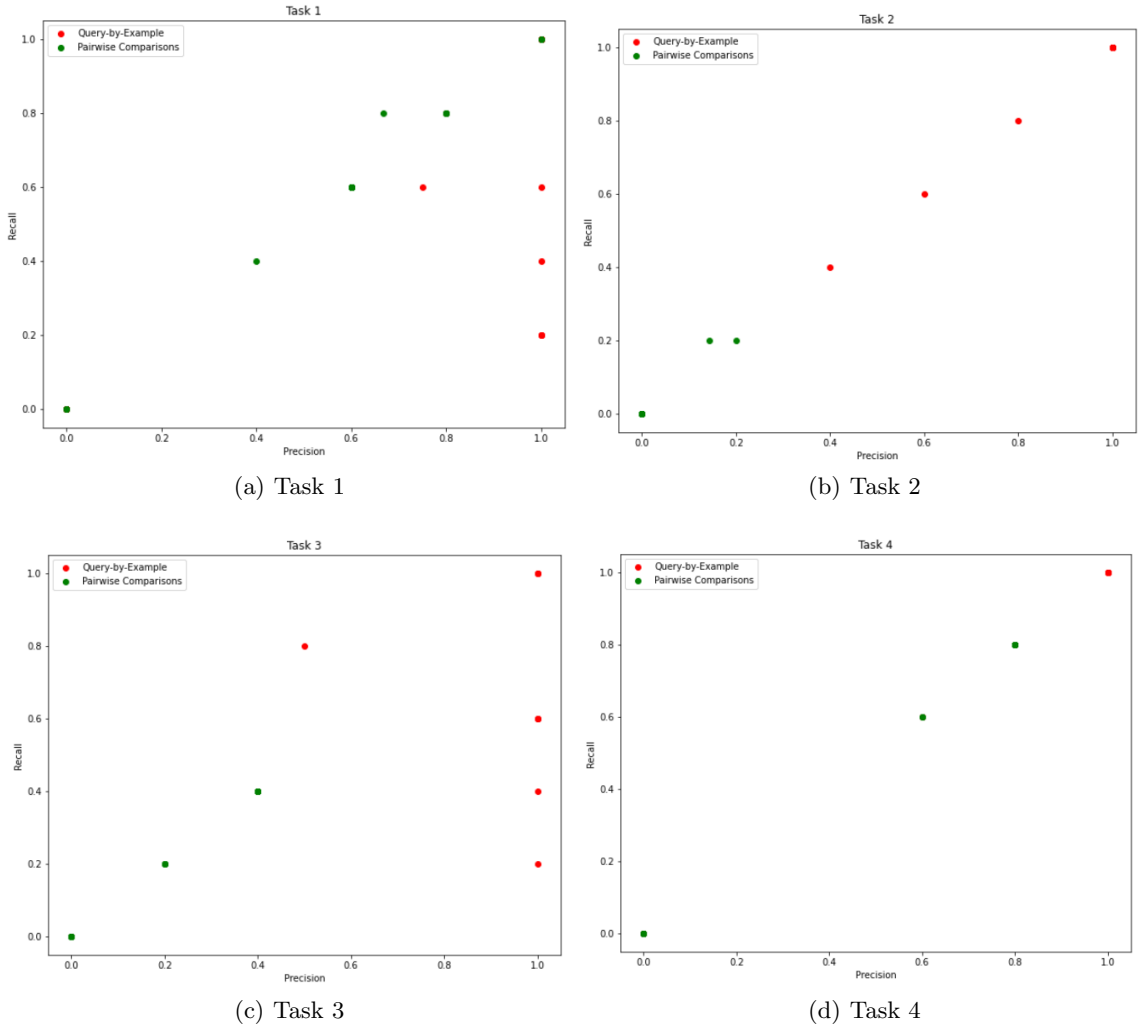
Usability Score vs. F-measure Figure 9 shows the relationship between Usability Score and F-measure obtained. For the case of QBE, a positive, albeit slight, correlation is observed for all four tasks. This suggests that participants who tend to perform well using this elicitation approach also tend to prefer it more. Conversely, there is no correlation (or a very slight negative one) in the case of PCs.

6 Discussion

Below is an assortment of the impressions recorded in the final part of the experiment, in which participants are asked to compare the two elicitation approaches:

C1 *Even though QBE does not always return results, it enables a simpler and more specific search.*
C2 *The PC approach requires greater knowledge about computers than the QBE approach.*
C3 *PC may be a better option for a user who is unsure of the technical specs required.*
C4 *QBE mimicks the process of purchasing a computer more realistically.*
C5 *With PC it is hard to understand which criterion is most important.*
C6 *Generally when buying a laptop I do not prioritise certain criteria over others, and I prefer to consider them separately.*

The above comments reflect the usability scores awarded to each elicitation approach. In general, the QBE approach is favoured particularly for the tasks where criteria values are explicit because it is easier to translate a technical specification to a QBE formulation (C4). However, the potential



(a) Task 1 (b) Task 2

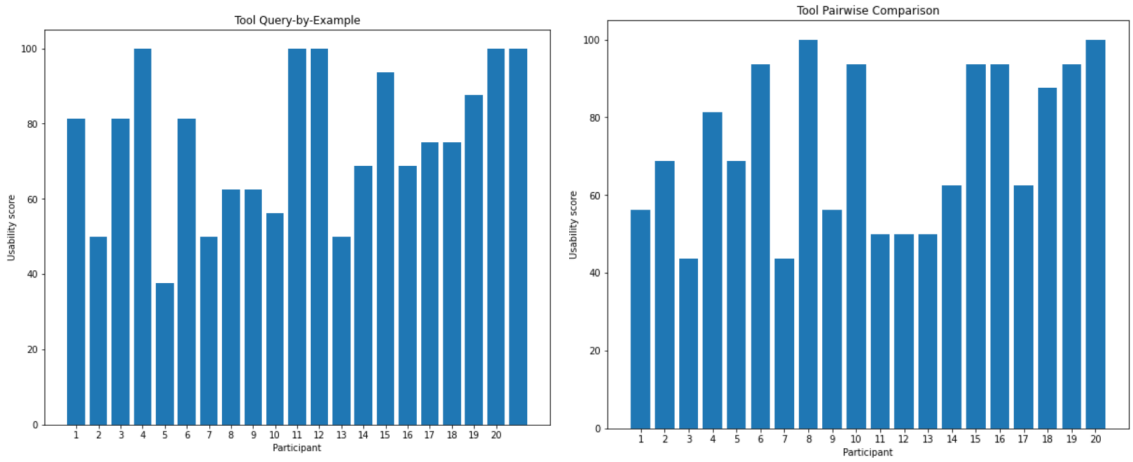
(c) Task 3 (d) Task 4

Fig. 5. Precision vs. Recall

drawback posed by tasks where criteria values are not stated explicitly are that QBE may result in a empty result set is noted (C1).

Indeed, all e-commerce sites known to the authors that sell technological items use a QBE-based (rather than PC-based) approach to search for results, so this is probably the most familiar way for the participants to select a laptop computer. We note that the QBE approach is often coupled with filters that suggest sensible ranges, to avoid the potential problem of an empty result set being returned.

The comments given about the PC approach are somewhat ambiguous. While one participant considers it may lend itself better to people with less knowledge about the technical specifications (*C3*), others consider that a deeper understanding is required (*C2,C5,C6*), possibly because the



(a) Query-by-example (b) Pairwise Comparisons

Fig. 6. Usability score (per person)

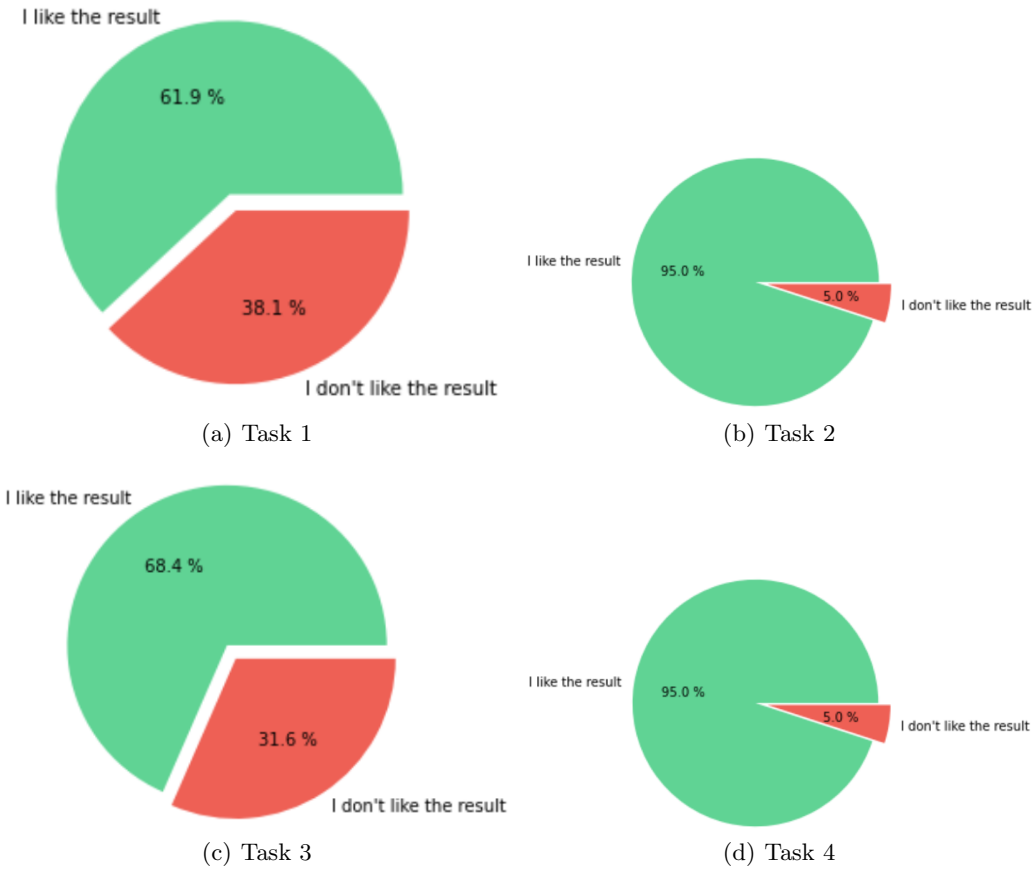
trade-offs that present themselves between different criteria (e.g., price vs. weight) may not always be apparent to less technically-minded people. The results in the previous section indicate that in cases where requirements are not precisely stated, i.e., Tasks 1 and 3 where the criteria values are implict rather than explicit, there is considerably higher satisfaction with the result sets obtained. As such, the PC elicitation approach may prove useful as a complementary elicitation approach to QBE for technology e-commerce websites.

7 Conclusions

In this paper, we compare the performance and usability of QBE and PC as preference elicitation techniques for search tools envisioned to be used on a e-commerce laptop computer sales websites. We find that while users do not directly express a preference for one tool over the other, for tasks where criteria values are not defined explicitly that PCs offers a promising preferance elicitation approach. We note that for the domain selected, i.e., laptop search on an e-commerce website, QBE is the habitual form of searching for results. As such, the results may be rather domain dependent, and fruitful future work may involve the exploration of usability of Pairwise Comparisons for other domains.

References

1. Abel, E., Galpin, I., Paton, N.W., Keane, J.A.: Pairwise comparisons or constrained optimization? a usability evaluation of techniques for eliciting decision priorities. International Transactions in Operational Research (2020)
2. Albaum, G.: The likert scale revisited. Market Research Society. Journal. **39**(2), 1–21 (1997)
3. Galpin, I., Abel, E., Paton, N.W.: Source selection languages: a usability evaluation. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. pp. 1–6 (2018)

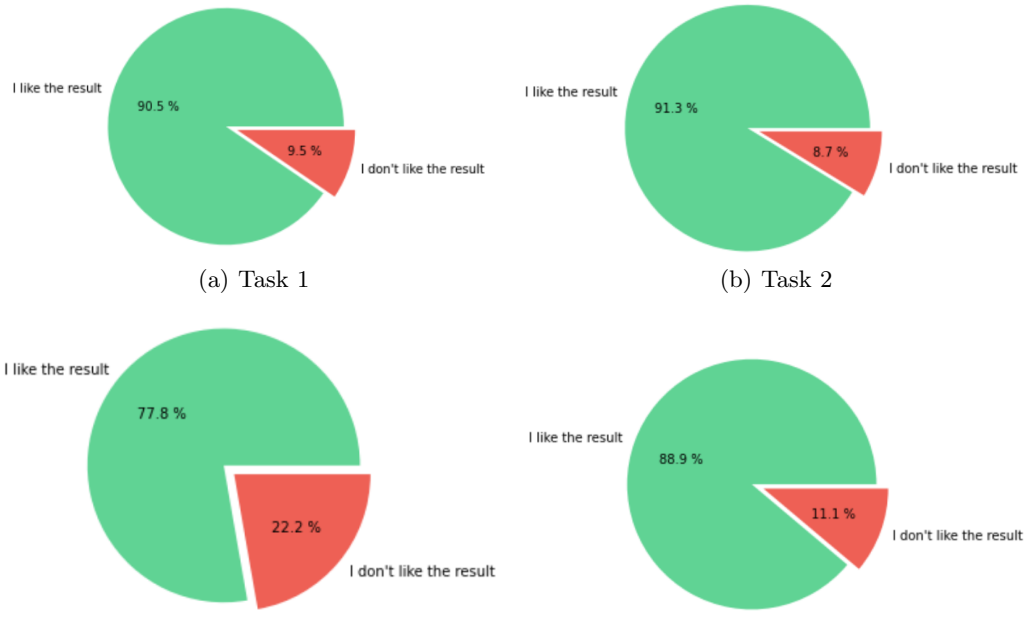


(a) Task 1 (b) Task 2

(c) Task 3 (d) Task 4

Fig. 7. Result set evaluation for Query-by-Example

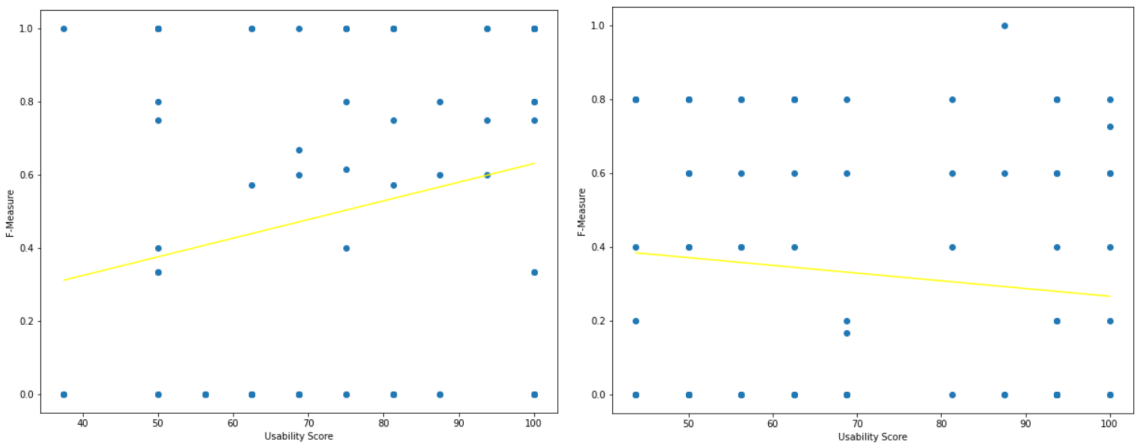
4. Kalloori, S., Ricci, F., Gennari, R.: Eliciting pairwise preferences in recommender systems. In: Proceedings of the 12th ACM Conference on Recommender Systems. pp. 329–337 (2018)
5. Leon-Medina, J.C., Galpin, I.: Refdatacleaner: A usable data cleaning tool. In: International Conference on Applied Informatics. pp. 102–115. Springer (2019)
6. Mander, B.: How mercadolibre emerged as an ecommerce titan: Latin america's answer to alibaba valued at $63bn as consumers turn to online shopping. Financial Times (Nov 2020), https://www.ft.com/content/446558a8-c0b2-449c-97a5-53b3956cd427
7. Saaty, T.L.: A scaling method for priorities in hierarchical structures. Journal of mathematical psychology **15**(3), 234–281 (1977)
8. Saaty, T.L.: Decision making for leaders: the analytic hierarchy process for decisions in a complex world. RWS publications (1990)
9. Sauro, J.: Measuring usability with the system usability scale (sus), https://measuringu.com/sus/, online; Accessed 10 May 2019
10. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to information retrieval, vol. 39. Cambridge University Press Cambridge (2008)
11. Sepliarskaia, A., Kiseleva, J., Radlinski, F., de Rijke, M.: Preference elicitation as an optimization problem. In: Proceedings of the 12th ACM Conference on Recommender Systems. pp. 172–180 (2018)



(a) Task 1 (b) Task 2

(c) Task 3 (d) Task 4

Fig. 8. Result set evaluation for Pairwise comparisons



(a) Query-by-Example ($\rho = 0.222$) (b) Pairwise Comparisons ($\rho = -0.125$)

Fig. 9. Correlation between Usability Score and F-measure

12. Thomas, J.C., Gould, J.D.: A psychological study of query by example. In: Proceedings of the May 19-22, 1975, National Computer Conference and Exposition. pp. 439–445 (1975)

13. Zloof, M.M.: Query by example. In: Proceedings of the May 19-22, 1975, National Computer Conference and Exposition. pp. 431–438 (1975)