Intl. Trans. in Op. Res. xx (2021) 1–17

Pairwise Comparisons or Constrained Optimization? A Usability Evaluation of Techniques for Eliciting Decision Priorities

Edward Abel[a,*], Ixent Galpin[b], Norman W. Paton[a] and John A. Keane[a]

[a]*Department of Computer Science, University of Manchester, Manchester M13 9PL, United Kingdom*
[b]*Facultad de Ciencias Naturales e Ingeniería, Universidad Jorge Tadeo Lozano, Bogotá, Colombia*
E-mail: edabelcs@gmail.com [Abel]; ixent@utadeo.edu.co [Galpin]; norman.paton@manchester.ac.uk [Paton]; john.keane@manchester.ac.uk [Keane]

Received DD MMMM YYYY; received in revised form DD MMMM YYYY; accepted DD MMMM YYYY

Abstract

Decision support methodologies provide notations for expressing and communicating the priorities that inform a decision. Whereas a substantial literature has explored the theoretical merits of such notations and methodologies, much less work has investigated their usability in practice, which is of vital importance for their widespread adoption by users. In this paper we explore the usability of two well-known preference elicitation techniques, *pairwise comparisons* and *constrained optimization*.

The techniques were explored through two contrasting crowd worker experiments, a preliminary one evaluating *recognition*, i.e., the ability to identify the most suitable formulation for a given task, and the other *synthesis*, i.e., the ability to construct formulations for a given task. The tasks are based on a case study involving source selection, a well-known problem in the data integration domain.

The results of the empirical evaluation show that, overall, *pairwise comparisons* resulted in significantly higher performance than *constrained optimization*, yet there is negligible difference between the usability appraisals for each technique. Furthermore, we observed that the technique that participants perform better with is not necessarily the one that they consider more usable.

Keywords: decision-making method; usability; pairwise comparison; constrained optimization; user preference elicitation; experimental evaluation

1. Introduction

Decision support methodologies provide notations for expressing and communicating the priorities that inform a decision. Whilst there is a substantial literature on notations and methodologies, there has been less work investigating their usability in practice. This is an important area of study as, for a decision

[*] Author to whom all correspondence should be addressed (e-mail: edabelcs@gmail.com).

methodology to be adopted *and* applicable in the real world, it is vital that user requirements are captured effectively. In this paper, building on our preliminary study (Galpin et al., 2018), we report on a usability evaluation using two well known preference elicitation techniques, specifically *pairwise comparisons* (PC), and *constrained optimization* (CO). Both techniques enable users to express their requirements as *formulations* in terms of domain-specific criteria. PC, employed in several Multiple-Criteria Decision Analysis (MCDA) methodologies such as the Analytic Hierarchy Process (Saaty, 1980), enables consideration of a pair of criteria at a time, to allow a user to define their preference, and strength of preference, between the pair helping to induce a separation of concerns (van Til et al., 2014). CO is inspired by classical linear programming (Dantzig, 1951); a single criterion is specified as an optimization goal, in the context of zero or more constraints cast as inequalities in terms of the available criteria.

We chose these techniques for eliciting and expressing user preferences, because they are two commonly-used and contrasting techniques. PC have been utilized within diverse applications (Vargas, 1990) and decision support software tools (Siraj et al., 2015), as has CO (Vanderbei, 2008; Mahmoud et al., 2010). Furthermore, CO is a quantitative or absolute approach, whereas PC is a qualitative or ratio-based approach. These techniques are therefore deemed to be interesting and contrasting points to explore in the user preference elicitation spectrum.

To understand the effectiveness and usability of the proposed techniques, we carried out an empirical evaluation using crowd workers to compare user performance, and opinions about the usability of each technique. In this way, we seek to explore the proposed techniques from a user perspective, as opposed to a more purely theoretical comparison of user preferences techniques, such as in (Barzilai, 2005).

We used *source selection*, the problem in data integration that addresses the identification of data sources that are most aligned to a user's needs, as our case study. In the context of source selection, a criterion refers to an aspect of data quality for a source, such as the validity or completeness of certain data items. For example, in the real-estate domain which we use here, the *location quality* criterion considers the availability of valid postcode and town data. Source selection is particularly relevant and challenging in Big Data and Data Science settings due to the huge number of potential data sources in certain domains, and plays an important role in improving data quality and reducing data integration costs (Lin et al., 2019). As such, the ability to find, access and make sense of data sets is vitally important for data scientists seeking to obtain data that will be most fit-for-purpose (Chapman et al., 2020). Recently, source selection has been explored as a multi-criteria problem in Rekatsinas et al. (2016) and Abel et al. (2018) without, as yet, consideration of the usability of preference elicitation. Our empirical evaluation comprises two experiments:

- Experiment 1 evaluates *recognition*, i.e., a participant's ability to identify, from a list of options, the most suitable formulation for a given task – analogous to a scenario in a Database Management System (DBMS) in which a user can choose from a set of predefined, *canned* queries. The results for this preliminary experiment are previously reported in Galpin et al. (2018), and complemented with further data analysis and findings in this paper.
- Experiment 2 evaluates *synthesis*, i.e., a participant's ability to construct formulations for a given task, analogous to a scenario in a DBMS in which *ad hoc*, interactive queries are posed by a user.

The contributions of this paper are (i) the proposal of a methodology to evaluate performance and usability of two well-known decision-making techniques, and (ii) the implementation of an empirical evaluation of the methodology, using source selection over data from the real-estate domain as a case study, that shows significant differences in performance for these techniques. This paper is structured as follows. Section 2 explores relevant literature; experiment design is described in Section 3; results are presented in Section 4; Section 5 concludes.

2. Literature Review

A previous study (Millet, 1997) compared elicitation techniques involving PC and direct estimation, comparing graphical, numerical and verbal representations to determine a set of weights and their usability as an elicitation technique. Our work differs as we compare CO, an approach that has no notion of criteria weights. Evaluation was done by comparing technique performance wrt. the ground truth and ease of use for the problem of determining the ratios between the areas of five two-dimensional shapes. Only a single task is used (so it could be argued that the task lends itself better to a particular approach) whereas we employ a diverse set of tasks. The experimentation found that participants obtained higher performance using PC. However, a general trade-off was observed between technique performance and perceived usability.

Sato (2004) compared results from citizen survey responses in which questions were posed as either multiple choice or PCs. Resulting sets of rankings, of the set of elements under consideration, were then compared for each technique. The analysis found that each technique yielded a different ranking of the elements under consideration, highlighting the potential for the method of elicitation to impact upon decision results. In the study, with purely subjective questions, the outcomes for the two techniques can simply be observed as being different. In contrast, our work looks to explore tasks for which we are able to calculate performance metrics, and thus be able to compare the performance and usabilty of different techniques.

Regarding the efficiently of PC, Carterette et al. (2008) found that for elicitation of user preferences, pertaining to items under consideration, PC preferences between items can be elicited more quickly compared to elicitation of absolute judgements of the items.

The number of preference relations defined by a user, in the form of pairwise comparisons, is explored via simulated experiments in Dong et al. (2019), to explore whether incomplete preference relations are better than complete preference relations, in terms of how well the different preference relations are at approximating the true priority vector. The experiments found that one approach did not always achieve the best performance. Indeed, the best approach was dependent on the rational degree of the decision maker when defining his/her pairwise comparisons which, if high enough, led to incomplete reference relations outperforming complete preference relations.

Usability studies have explored the relationship between a user's performance and his/her usability preferences. Nielsen and Levy (1994) analysed the relation between user preferences and performance, collating comparisons via meta-analysis of previous interface studies. Overall they found some positive correlation between participant performance and subjective preference regarding usability. However, they note that the timing of usability evaluations can impact an evaluation, and that opinions users express before trying a system can be less indicative of a user's eventual opinion after using the system.

Frøkjær et al. (2000) explored usability from the perspective of comparing the relationship between effectiveness, efficiency and satisfaction, through empirical experimentation of participants across a range of 20 information retrieval tasks. The experimentation showed little correlation between these three aspects and thus more independent facets are required to determine overall usability.

Analysis of performance and preferences of tasks being performed on both computer and mobile devices was explored by Adepu and Adler (2016). Here, they found that although task performance was significantly better on computers than on mobile devices, participants exhibited a preference to perform the tasks on mobile devices rather than computers.

Given that different multi-attribute decision-making (MADM) methods often produce different outcomes, Yeh (2002) explores the issue of choosing a MADM method via a data driven philosophy. This work proposes a validation approach for the selection of MADM methods that considers a given problem's data. The approach determines the most appropriate method as the one that best reflects the decision information content, taking into account information such as the relative contrast intensity of the alternatives' performance ratings on different attributes. Therefore, choosing a method is viewed as a data driven approach, without direct consideration of users. In contrast, our work presents a more user focused analysis, exploring how users perform with different elicitation techniques for the same tasks.

Some papers focus on the decision-making method procedure as a whole, rather than the elicitation approach, which is the focus of our study. An empirical user evaluation of goal programming, a variant of CO with support for multiple objectives is presented in Buchanan (1994). Users express requirements by ranking potential solutions, rather than specifying an explicit formulation, as in our experiments. They found that of the approaches to solution search evaluated, users prefer the unstructured approach, which gives them more control, despite taking longer than other approaches. A user study comparing various decision analysis and MCDM techniques is presented in Corner and Buchanan (1997); they conclude that the choice of method has no "order effect" and that decision makers should be screened individually to determine the best method for them. The usefulness of the MCDA methodology AHP is explored in Ishizaka et al. (2011) – which makes extensive use of PC for user preference elicitation – in helping decision-makers make a choice. Here, using incentivized empirical experimentation (Becker et al., 1964), they found that participants benefited from the ranking AHP produces. The combination of both PCs and multi choice goal programming was utilised by Kırış (2014) as separate stages, to create a hybrid decision model looking to leverage the attributes of each technique together. Within the approach the user makes use of both techniques as part of a sequential two-stage process, whereas our work looks to explore and compare how users fare using a single technique at a time.

In recommender systems – where the objective is for the system to provide "successful" recommendations in contrast to our experiments in which the user needs to complete specific tasks – studies have explored similar contrasting preference elicitation approaches, viz., that of PC of items and direct single item preference rating, with differing findings. Jones et al. (2011b) found that it is easier to decide which item is preferred among two, rather than rating them using some predefined scale. In contrast, Nobarany et al. (2012) concluded that "[single item] rating is the more familiar and less cognitively demanding form of judgement" in a setting where PC are explored as Boolean preference, rather than via a varying strength scale as in our experimentation. Similarly, Bottomley and Doyle (2013) explored eliciting "item" judgements via point allocation (whereby a fixed sum of 100 points are distributed among a set of criteria) and direct ratings (whereby rating are assigned on a fixed scale, such as 0–10) using objectively verifiable perceptual tasks. All criteria are considered together (point allocation) or one at a time (direct ratings) to seek sets of criteria weights. They observed that direct ranking is simpler to use and results in greater performance in terms of how accurately participant judgement matched the true results of the experiment. Our experiments differ in that they compare a ratio-based approach considering criteria in pairs against CO, a more contrasting approach.

Authors such as Kalloori et al. (2019) have explored approaches to facilitate elicitation of preferences as both absolute and relative PC, looking to exploit the metrics of both within a hybrid model. In this work, PC preferences are elicited first to predict a user's preference order of items, and subsequently, are fine-tuned by eliciting absolute user preferences.

3. Description of the Experiment

This section describes the recognition and synthesis experiments that we used to gain an insight into the effectiveness and usability of the investigated techniques. Both experiments collect evidence about user *performance*, i.e., effectiveness when employing a technique, as well as *usability questions*, to obtain user opinion about a technique.

Each experiment initially presents an introduction to the dataset used, and provides an explanation of the data quality criteria of interest. Subsequently, a short tutorial is presented for one of the techniques, and four source selection tasks are given to the user, which must be solved using that technique. After completing the tasks, the user is asked four multiple-choice questions about his/her opinion of the technique, adapted from the well-established System Usability Scale (SUS) questionnaire (Brooke et al., 1996).

3.1. Tasks

The types of tasks selected for the experiment are intended to be varied and have diverse characteristics, in order to cover a broad range of possible user requirements. Table 1 describes the tasks in generic, domain-independent terms, alongside an example task from the real estate domain. Given that the techniques have different functionalities, having diverse tasks enables usability to be evaluated across cases such as when one technique is, in principle, more suitable for a task than the other.

For each task, the user is presented with a natural language description of the data to be obtained, and is required to either recognise or synthesize (depending on the experiment), using the current technique, the most effective formulation, i.e., the one that will obtain the data with the quality characteristics required from the task description. For the **Recognition Experiment**, each performance question presents four possible formulations from which the user may choose, as shown for PC in Figure 1a and for CO in Figure 1b. These formulations are selected such that they yield result sets of starkly contrasting quality, and the user needs to select the formulation that will retrieve the data that best matches the task description. Given the small set of possible answers, only one attempt is permitted for each task.

For the **Synthesis Experiment**, performance questions present an interface for users to construct formulations from scratch. The interfaces for PC and CO are shown in Figures 2a and 2b respectively. For example, for PC, the user may add one or more PC, selecting criteria and their relative importance, for each comparison, from drop-down menus. After constructing the formulation, the corresponding result set is shown. Based on this, the user may then decide to revise the formulation. Mimicking a real-world scenario in which a user may come up with an initial formulation and make subsequent adjustments, the user is permitted multiple attempts until he/she is satisfied with the result.

Task	Abstract Task	Concrete Task
1	*Obtain data that has a certain quality threshold for a single criterion.*	You have a property dataset with poor location data. However, you need to obtain data with the best location quality possible. Ideally, you should obtain as many records as possible that have some location information (even if it is just the street name).
2	*Obtain data that has a certain quality threshold for two given criteria.*	Due to falling sales of properties, you need to compare price information between properties with a similar location. Preferably, you should aim to get both town and street location data. If that is impossible, just the town will do.
3	*Obtain data that is of high quality for at least one of two conflicting criteria.*	You have a property dataset in which most records have high quality data for only one quality measure. For example, if a record has high quality price data, it is unlikely to have high quality location or room information. You wish to obtain records that have high quality data for either location or room information. You are not worried about price data.
4	*Obtain data that is of high quality for one criterion, but of low quality for another criterion.*	You have a property dataset, and wish to retrieve records with good price information; however, you do not wish to retrieve any records with complete location data.

Table 1
Generic task descriptions with the corresponding domain-specific task.

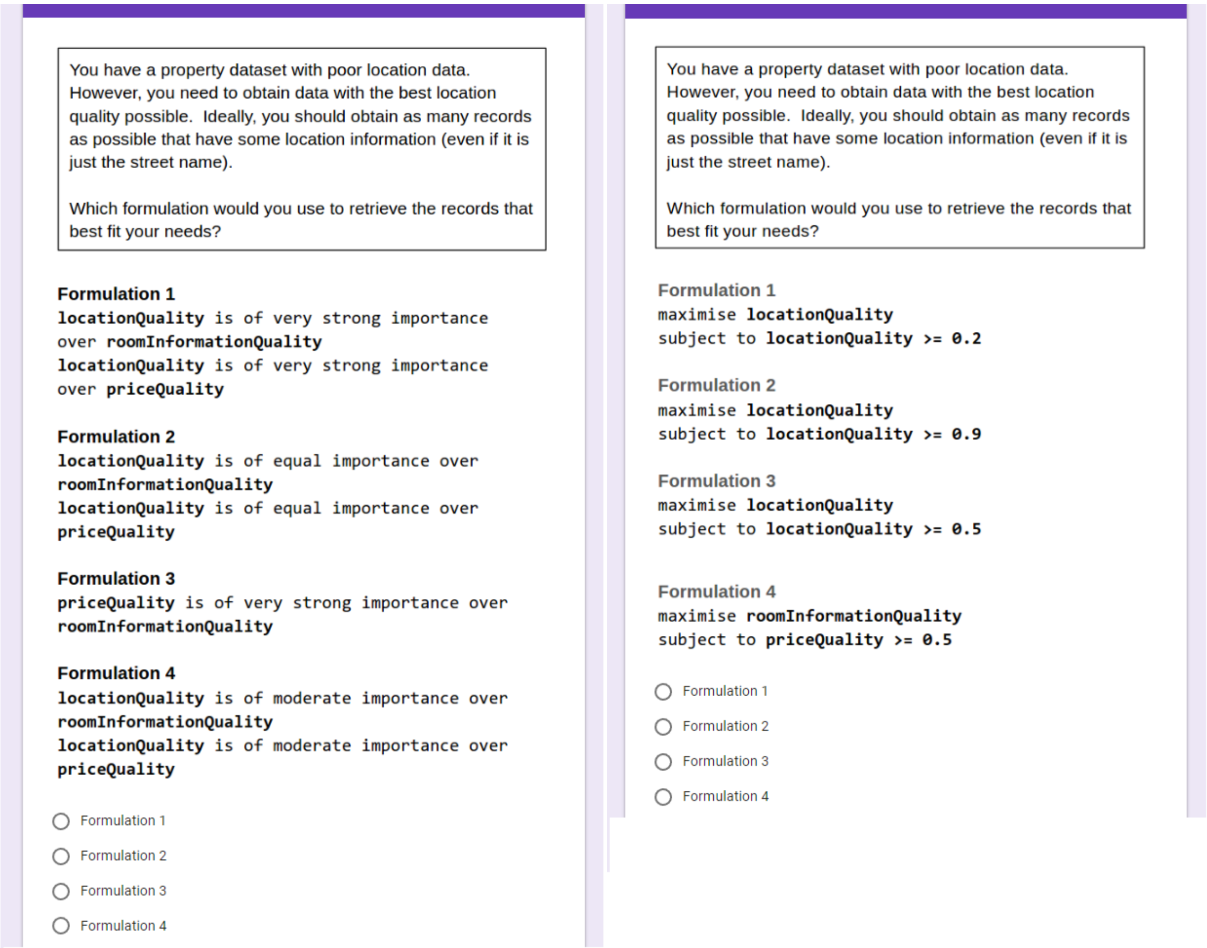
3.2. Evaluation Measures

The experiments employ three types of evaluation measure:

- *User performance* measures the overall effectiveness of a user at recognising/synthesising the most effective formulation for a task using a given technique.
- *User efficiency* measures consider the time taken and the number of attempts taken for each task, and are only recorded for the synthesis experiment due to its interactive nature.
- The *usability score* aims to capture subjective user preference about each technique. It is a value between 0 and 100, and is calculated in a similar manner to SUS (Brooke et al., 1996).

For both experiments, *User Performance* is related to the F-measure of the result set obtained for a given formulation. We use F-measure as it is a widely utilised measure within information retrieval that provides a measure of accuracy considering both precision and recall, as well as a succinct means to compare result sets that may be of different sizes. The F-measure for a given result set $\mathbf{R}$ with respect to a task T and formulation f, denoted $F(\mathbf{R}_f^T)$, is calculated thus:

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



You have a property dataset with poor location data. However, you need to obtain data with the best location quality possible. Ideally, you should obtain as many records as possible that have some location information (even if it is just the street name).

Which formulation would you use to retrieve the records that best fit your needs?

Formulation 1

```
locationQuality is of very strong importance
over roomInformationQuality
locationQuality is of very strong importance
over priceQuality
```

Formulation 2

```
locationQuality is of equal importance over
roomInformationQuality
locationQuality is of equal importance over
priceQuality
```

Formulation 3

```
priceQuality is of very strong importance over
roomInformationQuality
```

Formulation 4

```
locationQuality is of moderate importance over
roomInformationQuality
locationQuality is of moderate importance over
priceQuality
```

○ Formulation 1

○ Formulation 2

○ Formulation 3

○ Formulation 4

(a) Pairwise Comparisons

You have a property dataset with poor location data. However, you need to obtain data with the best location quality possible. Ideally, you should obtain as many records as possible that have some location information (even if it is just the street name).

Which formulation would you use to retrieve the records that best fit your needs?

Formulation 1

```
maximise locationQuality
subject to locationQuality >= 0.2
```

Formulation 2

```
maximise locationQuality
subject to locationQuality >= 0.9
```

Formulation 3

```
maximise locationQuality
subject to locationQuality >= 0.5
```

Formulation 4

```
maximise roomInformationQuality
subject to priceQuality >= 0.5
```

○ Formulation 1

○ Formulation 2

○ Formulation 3

○ Formulation 4

(b) Constrained Optimization

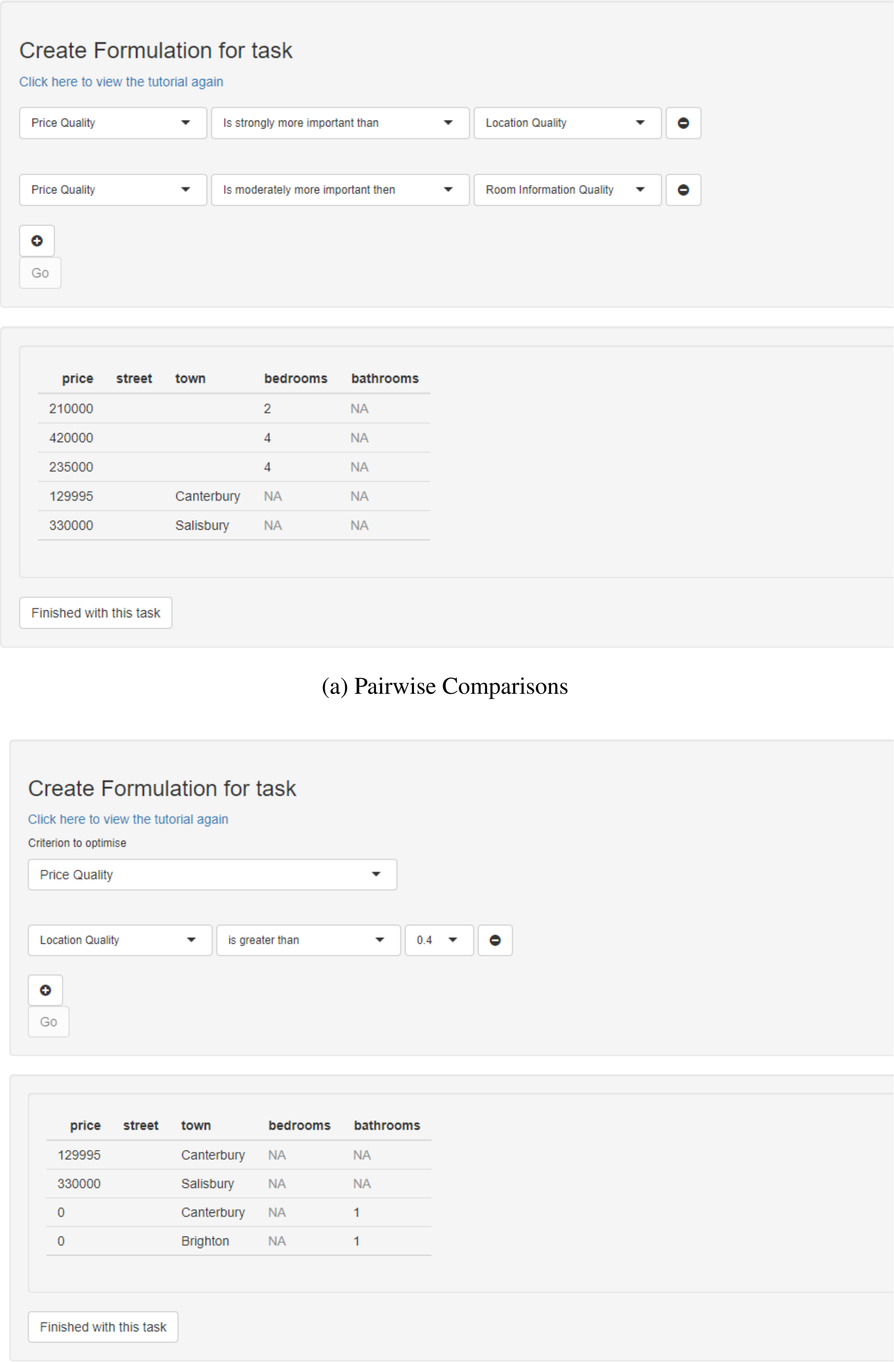
Fig. 1. Recognition screen shots for *Task 1*.

where $\mathbf{R}_f^T$ is result set obtained using formulation f and

$$\mathsf{Precision}(\mathbf{R}_f^T) = \frac{|TP_f^T|}{|TP_f^T \cup FP_f^T|} \tag{2}$$

and

$$\mathsf{Recall}(\mathbf{R}_f^T) = \frac{|TP_f^T|}{|TP_f^T \cup FN_f^T|} \quad . \tag{3}$$



Create Formulation for task

Click here to view the tutorial again

Price Quality | Is strongly more important than | Location Quality

Price Quality | Is moderately more important then | Room Information Quality

price	street	town	bedrooms	bathrooms
210000			2	NA
420000			4	NA
235000			4	NA
129995		Canterbury	NA	NA
330000		Salisbury	NA	NA

(a) Pairwise Comparisons

Create Formulation for task

Click here to view the tutorial again

Criterion to optimise

Price Quality

Location Quality | is greater than | 0.4

price	street	town	bedrooms	bathrooms
129995		Canterbury	NA	NA
330000		Salisbury	NA	NA
0		Canterbury	NA	1
0		Brighton	NA	1

(b) Constrained Optimization

Fig. 2. Synthesis screen shots for *Task 1*.

The set of true positives, denoted TP_f^T, is defined as follows:

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

where $\mathbf{F}_T$ is the set of true positives, i.e., the records which are deemed to be fit-for-purpose according to the natural language description for task T. For example, when seeking data to carry out a price comparison task, a true positive would be a record for which complete price information is present. The sets of false positives and false negatives are similarly defined:

$$FP_f^T = \{r | r \in \mathbf{R}_f \wedge r \notin \mathbf{F}_T\} \tag{5}$$

$$FN_f^T = \{r | r \notin \mathbf{R}_f \wedge r \in \mathbf{F}_T\} \tag{6}$$

It is important to note that the performance measures differ slightly for the two experiments. For the **Recognition Experiment**, the performance is based on the notion of *Answer Rank*, a discrete scale of finite values. Each set of four formulations for a task results in distinct F-measures, from which an ordinal ranking can be determined, resulting in an answer rank of between 1 (best answer) and 4 (worst answer).

For the **Synthesis Experiment**, performance maps to F-measure directly, meaning that it can be measured using a continuous scale from 0 to 1.

3.3. Experiment Setup

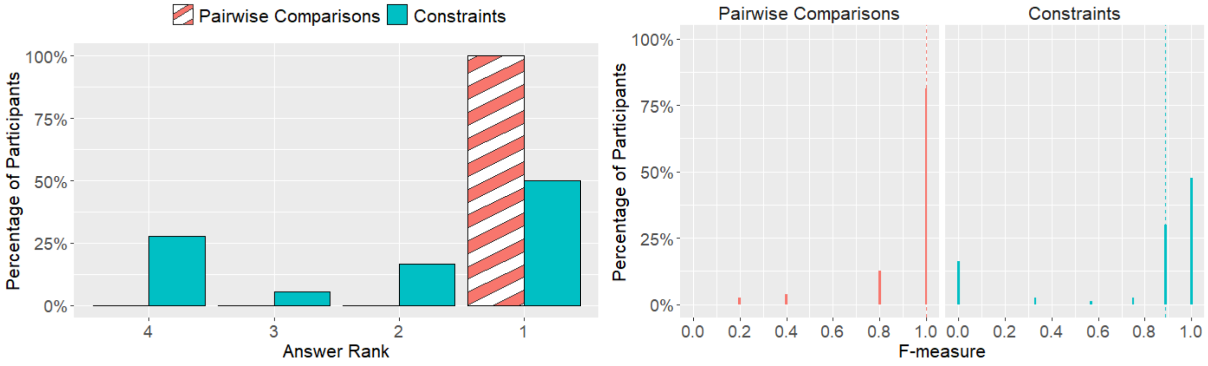
For our experiments, we curated a real-world dataset consisting of web-scraped data from the real-estate property domain, extracted via the DIADEM system by Furche et al. (2014). For the recognition experiment, we created a questionnaire using Google Forms, as shown in Figure 1. For the synthesis experiment, a bespoke web interface application using Shiny R was created, as shown in Figure 2. For both experiments, the four concrete tasks presented in Table 1 are given once for each technique, resulting in eight tasks in total. Two variations of each experiment were created, in which the techniques are presented in a different order. Half the users tackled PC first; the other half CO first. The task order (which is intended to be of increasing difficulty) did not vary. As such, each user explored the sequence of four tasks in the same order so that there was no variance introduced such as performance differences due to differing task orders, outside of our intended sequence of increasing difficulty.

For both experiments, participants were recruited using the Amazon Mechanical Turk[1] crowdsourcing platform. A Human Intelligence Task (HIT) was set up within Amazon Turk as an external link to the experiments. Participants were paid the UK minimum wage pro-rata according to the anticipated HIT duration time[2]. We prevented people from doing the HIT more than once by creating a custom qualification in Amazon Turk, which was awarded to each participant on completion of the HIT.

It was important to ensure that the participants demonstrated a high degree of understanding of the tasks as well as engagement. This was necessary as it would facilitate responses of high-quality regarding reliability of comprehension, as opposed to responses from a participant who did not sufficiently

[1] https://www.mturk.com/

[2] The UK minimum wage for 2018 is 7.83 GBP/hour.



(a) Task 1 recognition. $\tilde{A}_{PC} = 1$; $\tilde{A}_{CO} = 1.5$; *p*=0.00775.

(b) Task 1 synthesis (max of first three attempts). $\tilde{F}_{PC} = 1$; $\tilde{F}_{CO} = 0.889$; *p*=0.000445.

Fig. 3. User performance for each technique, Task 1.

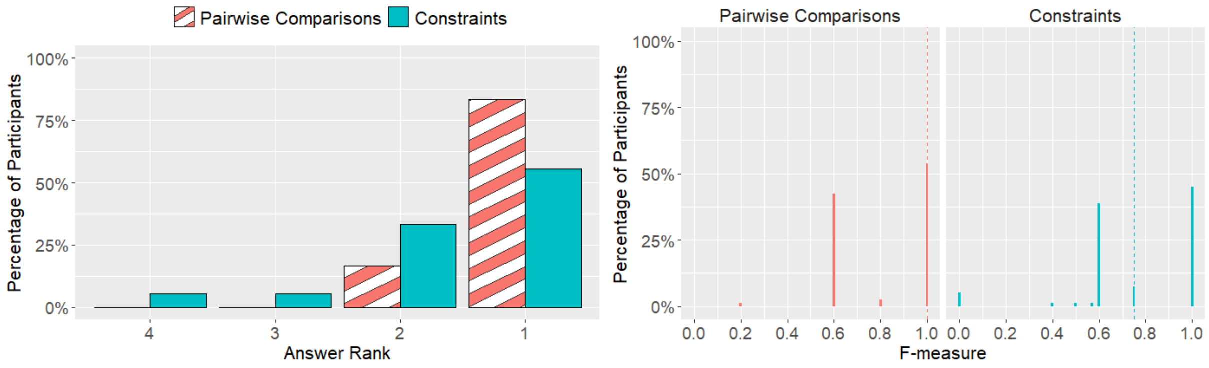
comprehend the tasks. We sought to ensure that we obtained high-quality responses through the use of validation questions, that enabled us to gauge participants' level of comprehension. To this end, we use additional task validation questions entailing (i) for recognition, presenting four result sets, from which the user is asked to identify one that most satisfies a task description and (ii) for synthesis, assessment of the quality of the result set obtained from the formulation they constructed. In this way, we can ascertain from users' comprehension of the tasks the reliability of their answers and keep for analysis only participants who demonstrated a high degree of understanding. From this, we obtained responses from 18 participants for recognition, and 80 participants for synthesis, in both cases with equal numbers of participants for both technique orders.

4. Results

4.1. User Performance

Figures 3-6 show the separate performance results for each of *Tasks 1-4* respectively. Figure 7 presents the combined user performance measures, for all four tasks, in each of the experiments. For synthesis, we show the maximum F-measure obtained during the first three attempts for a task, to attenuate the improvement in results that may be attributed to trial and error after many attempts. $\tilde{A}_T$ and $\tilde{F}_T$ denote the median answer rank and F-measure, respectively, for a technique $T \in \{PC, CO\}$. We employed paired, two tailed, Wilcoxon tests to determine whether there is a significant difference between the performance measures for the two techniques, and the corresponding *p*-values are shown in the captions.

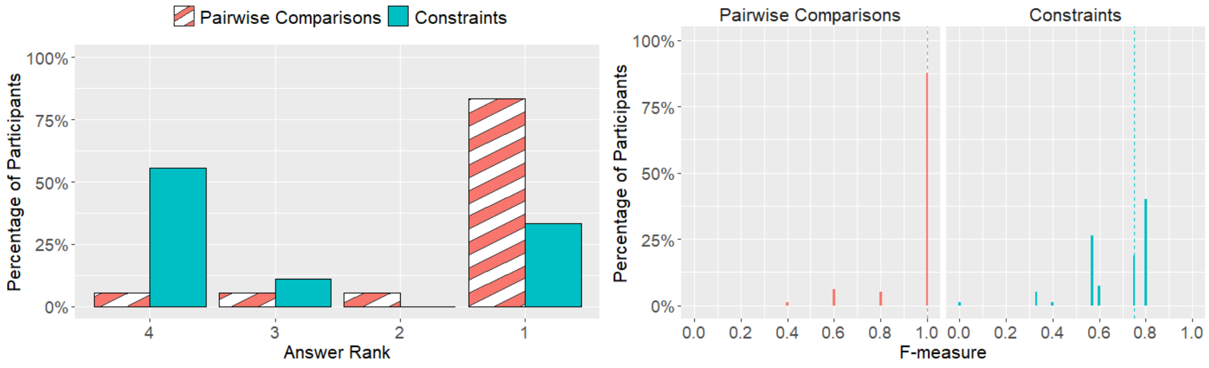
Task Level: For recognition in the case of *Task 1*, represented by Figure 3a, we observe that all participants chose the top-ranked formulation for PC. However, for CO the best formulation was not chosen as often, and selection of all four formulations occurs. The median answer rank is 0.5



(a) Task 2 recognition. $\tilde{A}_{PC} = 1$; $\tilde{A}_{CO} = 1$; p=0.081.

(b) Task 2 synthesis (max of first three attempts). $\tilde{F}_{PC} = 1$; $\tilde{F}_{CO} = 0.75$; p=0.0896.

Fig. 4. User performance for each technique, Task 2.



(a) Task 3 recognition. $\tilde{A}_{PC} = 1$; $\tilde{A}_{CO} = 4$; p=0.0003.

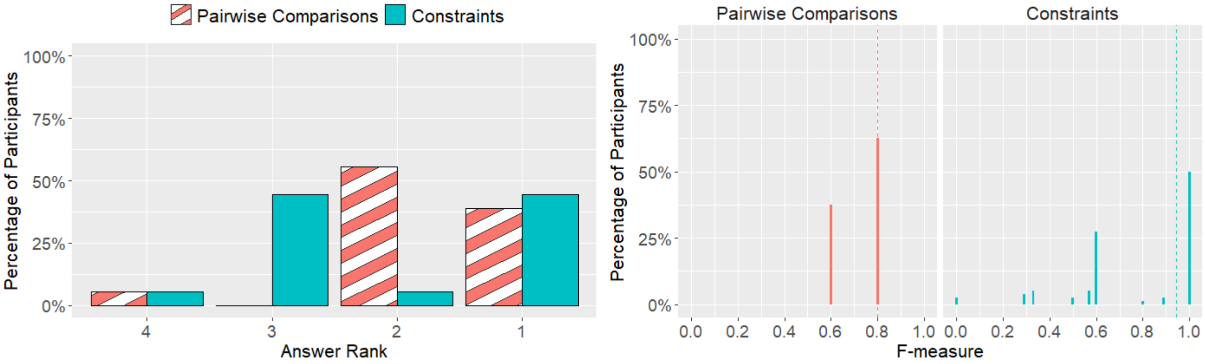
(b) Task 3 synthesis (max of first three attempts). $\tilde{F}_{PC} = 1$; $\tilde{F}_{CO} = 0.75$; p=$2.73e - 13$.

Fig. 5. User performance for each technique, Task 3.

rank positions higher for PC than for CO. For synthesis, the results show a similar trend: the median F-measure for PC is 0.111 higher for PC than for CO. The results are significant for both experiments.

In the case of *Task 2*, shown in Figure 4, the median answer rank is the same for both techniques in the recognition experiment. For synthesis, the median F-measure is 0.25 higher for PC than for CO. We note, however, that the results are not significant for either experiment.

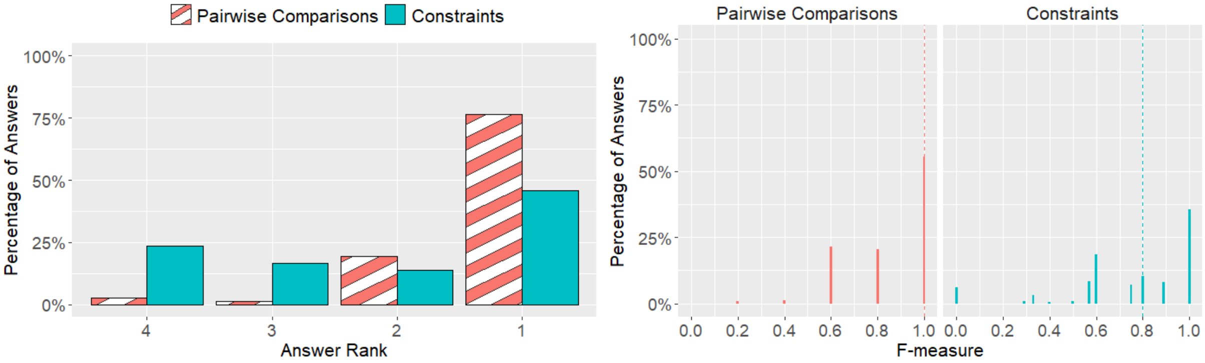
The difference in performance is starkest and most significant for *Task 3*. For recognition, as shown in Figure 5a, over 50% of users chose the *Rank 4* formulation with CO, resulting in the median answer rank being 2 rank positions higher for PC. For synthesis, as shown in Figure 5a, the highest F-measure obtained for CO is 0.8, indicating that the task lends itself better to PC, and $\tilde{F}_{PC}$ is 0.25 higher than $\tilde{F}_{CO}$.



(a) Task 4 recognition. $\tilde{A}_{PC} = 2$; $\tilde{A}_{CO} = 2.5$; p=0.115.

(b) Task 4 synthesis (max of first three attempts). $\tilde{F}_{PC} = 0.8$; $\tilde{F}_{CO} = 0.944$; p=0.720.

Fig. 6. User performance for each technique, Task 4.



(a) Recognition. $\tilde{A}_{PC} = 1$; $\tilde{A}_{CO} = 2$; p=$1.05e-6$.

(b) Synthesis (max of first three attempts). $\tilde{F}_{PC} = 1$; $\tilde{F}_{CO} = 0.8$; p=$3.29e-9$.

Fig. 7. User Performance for each technique; all tasks combined.

In contrast, for *Task 4* the results deviate from the general trend. For recognition, shown in Figure 6a, we observe that it is the only task in which the number of *Rank 1* answers given for CO is greater than PC. However, we still observe that selection of all four formulations occurs for CO. Although the median answer rank for PC remains higher than for CO, the difference in performance is not significant. For synthesis, shown in Figure 6b, although $\tilde{F}_{CO}$ is slightly higher than $\tilde{F}_{PC}$, the difference in performance is not significant either. It is interesting to note that, although *Task 4* lends itself better to CO, as PC does not include a mechanism for exclusion of data with certain characteristics, user performance with CO is not significantly better, as would be expected.

All Tasks Combined: Figure 7a shows the results grouped by *Answer Rank* for recognition. For PC the median answer rank across all tasks ranged between 1 and 3 for all users, and $\tilde{A}_{PC}$ was 1, whereas for

CO users the median answer rank across all tasks ranged between 1 and 4, and $\tilde{A}_{CO}$ was only 2. The value for $\tilde{A}_{PC}$ is 1 rank position higher than for $\tilde{A}_{CO}$. The difference in $\tilde{A}$ across both techniques is significant (p=$1.05e-6$).

The results for the synthesis experiment in Figure 7b show a similar, albeit less pronounced, pattern. For PC all users obtained an F-measure between 0.2 and 1, with $\tilde{F}_{PC}$ being 1. Performance was lower for CO: for this technique, the F-measures ranged from 0 to 1, and $\tilde{F}_{CO}$ was 0.8. The value for $\tilde{F}_{PC}$ is 0.2 higher than for $\tilde{F}_{CO}$. Again, these results are extremely significant (p=$1.69e-14$), even if the performance difference is lower than for recognition. Note that the lower performance of PC for *Task 4* leads to a net reduction in the overall performance difference.

Observation 1: Overall, higher performance with PC was observed, with high significance, for both experiments.

Observation 2: The results for *Tasks 4* show that CO underperforms even for a task for which it seems better suited.

4.2. User Efficiency

User efficiency measures were recorded for the synthesis experiment only, by carrying out detailed logging of user activity within the web application. Figure 8a presents a box plot showing distribution of times taken per task, for each technique. To aid visualization, outliers higher than 250 seconds are not shown in the plot. The plot shows that the median times taken per task for PC and CO were 72.5 s and 76.5 s respectively – a 3 s difference. The difference in times is significant (p=0.00933). Figure 8b shows the distribution of number of attempts taken per task for each technique. To aid visualization, outliers higher than 20 attempts are not shown in the plot. For this measure, the median was 2 for PC, and 3 for CO. The difference in number of attempts is significant (p=9.37e-08).

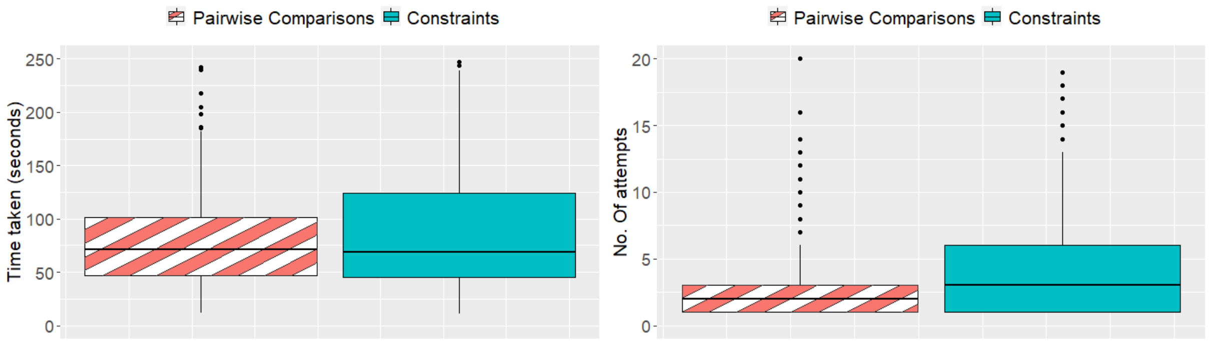
Observation 3: PC outperforms CO for the time taken and the number of attempt measures.

4.3. Usability Score

Figure 9 presents box plots relating to the usability scores obtained for each technique. Even though, for recognition, we observe a higher median usability score for PC, this is not significant (p=0.551). For synthesis, the difference in usability score is not significant either (p=0.998).

Observation 4: There is no significant difference observed between the usability scores of each technique.

4.4. Usability against Performance



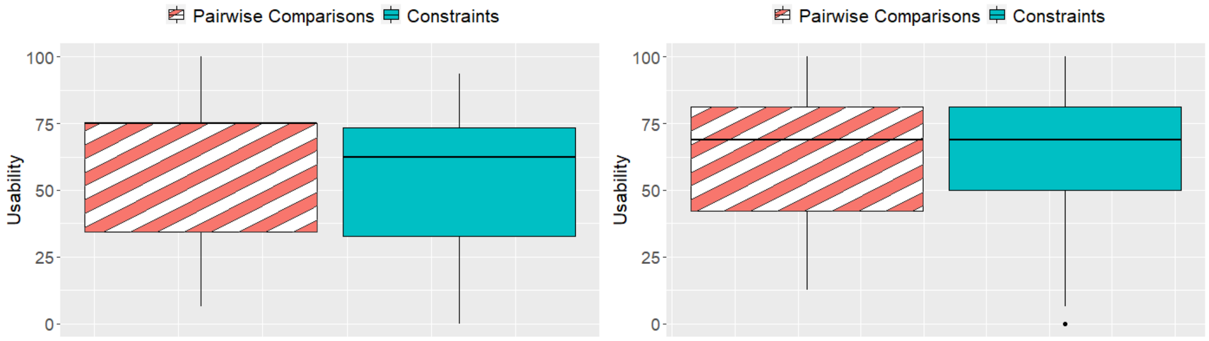
(a) Synthesis Time Taken (seconds)

(b) Synthesis no. of Attempts All Tasks

Fig. 8. User Efficiency

Comparing the relationship between participant performance and usability preference, we can see if there is any correlation between our participants performance and usability scores. Table 2 shows the correlation coefficients, and corresponding significance values, pertaining to performance vs usability, with respect to each technique, for the different experiments. From Table 2 we observe that apart from CO in the recognition experiment all the correlation coefficients are not significant (and are 0.2 or less), suggesting that we see little correlation between our participant's performance and his/her subjective usability preference. For CO in the recognition experiment we observe a correlation coefficient value of 0.45 suggesting a slight correlation here between performance and usability. However, this value is not significant (*p*-value of 0.06).

We further analyse the relationship between performance and usability preference between techniques. We calculated the percentage of users that fall into each of a possible nine categories and collate them



(a) Recognition

(b) Synthesis

Fig. 9. Usability Score.

	Pearson Correlation	Significance (p-value)
Pairwise Comparison Recognition	-0.20	0.424
Pairwise Comparison Synthesis	-0.15	0.197
Constraints Recognition	0.45	0.064
Constraints Synthesis	0.12	0.274

Table 2

Performance vs. Usability Correlation Coefficient

	Preferred PCCV	No preference	Preferred SOCO	Total
Better at PCCV	38.89%	11.11%	33.33%	83.33%
Same Performance	0.00%	0.00%	0.00%	0.00%
Better at SOCO	11.11%	0.00%	5.56%	16.67%
Total	50.00%	11.11%	38.89%	100%

Table 3

Recognition Confusion Matrix

	Preferred PCCV	No preference	Preferred SOCO	Total
Better at PCCV	33.75%	10.00%	28.75%	72.50%
Same Performance	2.50%	2.50%	3.75%	8.75%
Better at SOCO	6.25%	2.50%	10.00%	18.75%
Total	42.50%	15.00%	42.50%	100%

Table 4

Synthesis Confusion Matrix

into a 3 $\times$ 3 confusion matrix, shown in Tables 3 and 4. Here rows denote either better performance with PC, same performance for each technique, or better performance with CO, respectively. Similarly, columns denote either higher usability for PC, same usability, or higher usability for CO, respectively. The matrices for both experiments show similar patterns. For example, we observe that about 45% and 35% of users, for recognition and synthesis respectively, preferred a technique at odds with the technique with which they performed best. Furthermore, of the participants that performed best with PC, less than half, for both experiments, also preferred PC.

Observation 5: There is no relationship between a user's performance with a technique, and his/her's appraisal of the technique's usability.

Observation 6: We observe a significant proportion of users who prefer a different technique to the one with which they perform best.

5. Conclusions

Decision support methodology notations have been the subject of substantial literature focusing on their theoretical merits. In contrast, there has been much less work investigating their usability in practice, an important consideration for actual adoption by users. This paper explores the usability of two well-known techniques, PC and CO. We recruited participants using crowd sourcing to evaluate performance with each technique, as well as subjective user preference. The participants carried out tasks relating to source selection, a well-established problem in the data integration domain.

The empirical evaluation comprised two experiments with 18 and 80 participants respectively, with the following statistically significant results: (1) PC resulted in higher performance than CO, even for tasks for which it seems less well suited; (2) users required more time and more attempts for synthesis with CO than PC; (3) there is little, if any, difference between the usability appraisals for each technique; (4) there is a surprising mismatch between performance and subjective user preference.

In future work we will explore whether combining features of both techniques could create a hybrid that enables users to express requirements more effectively, when compared to separately.

Acknowledgement(s)

Funding: This work is supported by the VADA Programme Grant of the UK EPSRC [grant number EP/M025268/1].

References

Abel, E., Keane, J.A., Paton, N.W., Fernandes, A.A.A., Koehler, M., Konstantinou, N., Ríos, J.C.C., Azuan, N.A., Embury, S.M., 2018. User driven multi-criteria source selection. *Inf. Sci.* 430, 179–199.

Adepu, S., Adler, R.F., 2016. A comparison of performance and preference on mobile devices vs. desktop computers. In *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, pp. 1–7.

Barzilai, J., 2005. Measurement and preference function modelling. *International Transactions in Operational Research* 12, 2, 173–183.

Becker, G.M., DeGroot, M.H., Marschak, J., 1964. Measuring utility by a single-response sequential method. *Behav. Sci.* 9, 3, 226–232.

Bottomley, P.A., Doyle, J.R., 2013. Comparing the validity of numerical judgements elicited by direct rating and point allocation: Insights from objectively verifiable perceptual tasks. *Eur. J. Oper. Res.* 228, 1, 148–157.

Brooke, J., et al., 1996. SUS-A quick and dirty usability scale. *Usability Eval. Ind.* 189, 194, 4–7.

Buchanan, J.T., 1994. An experimental evaluation of interactive MCDM methods and the decision making process. *J. Oper. Res. Soc.* 45, 9, 1050–1059.

Carterette, B., Bennett, P.N., Chickering, D.M., Dumais, S.T., 2008. Here or there. In *European Conference on Information Retrieval*, Springer, pp. 16–27.

Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P., 2020. Dataset search: a survey. *The VLDB Journal* 29, 1, 251–272.

Corner, J.L., Buchanan, J.T., 1997. Capturing decision maker preference: Experimental comparison of decision analysis and MCDM techniques. *Eur. J. Oper. Res.* 98, 1, 85–97.

Dantzig, G.B., 1951. *Maximization of a linear function of variables subject to linear inequalities*. Wiley.

Dong, Y., Liu, W., Chiclana, F., Kou, G., Herrera-Viedma, E., 2019. Are incomplete and self-confident preference relations better in multicriteria decision making? A simulation-based investigation. *Inf. Sci.* 492, 40–57.

Frøkjær, E., Hertzum, M., Hornbæk, K., 2000. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 345–352.

Furche, T., Gottlob, G., Grasso, G., Guo, X., Orsi, G., Schallhart, C., Wang, C., 2014. DIADEM: Thousands of Websites to a Single Database. *PVLDB* 7, 14.

Galpin, I., Abel, E., Paton, N.W., 2018. Source selection languages: A usability evaluation. In *Proc. HILDA*, ACM, p. 8.

Ishizaka, A., Balkenborg, D., Kaplan, T., 2011. Does AHP help us make a choice? An experimental evaluation. *J. Oper. Res. Soc.* 62, 10, 1801–1812.

Jones, N., Brun, A., Boyer, A., 2011a. Comparisons instead of ratings: Towards more stable preferences. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1, IEEE, pp. 451–456.

Jones, N., Brun, A., Boyer, A., Hamad, A., 2011b. An exploratory work in using comparisons instead of ratings. In *EC-Web*, Springer, pp. 184–195.

Kalloori, S., Li, T., Ricci, F., 2019. Item recommendation by combining relative and absolute feedback data. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 933–936.

Kırış, Ş., 2014. Ahp and multichoice goal programming integration for course planning. *International Transactions in Operational Research* 21, 5, 819–833.

Lin, Y., Wang, H., Li, J., Gao, H., 2019. Data source selection for information integration in big data era. *Inf. Sci.* 479, 197–213.

Mahmoud, H.B., Ketata, R., Romdhane, T.B., Ahmed, S.B., 2010. Piloting a quality management system for study case using multi-choice goal programming. In *2010 IEEE International Conference on Systems, Man and Cybernetics*, IEEE, pp. 2500–2505.

Millet, I., 1997. The effectiveness of alternative preference elicitation methods in the analytic hierarchy process. *J. Multi-Criteria Decis. Anal.* 6, 1, 41–51.

Nielsen, J., Levy, J., 1994. Measuring usability: preference vs. performance. *Communications of the ACM* 37, 4, 66–75.

Nobarany, S., Oram, L., Rajendran, V.K., Chen, C.H., McGrenere, J., Munzner, T., 2012. The design space of opinion measurement interfaces: exploring recall support for rating and ranking. In *Proc. ACM SIGCHI*, ACM, pp. 2035–2044.

Rekatsinas, T., Deshpande, A., Dong, X.L., Getoor, L., Srivastava, D., 2016. Sourcesight: enabling effective source selection. In *ACM SIGMOD*, ACM, pp. 2157–2160.

Saaty, T.L., 1980. *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. McGraw-Hill.

Sato, Y., 2004. Comparison between multiple-choice and analytic hierarchy process: measuring human perception. *International Transactions in Operational Research* 11, 1, 77–86.

Siraj, S., Mikhailov, L., Keane, J.A., 2015. Priest: an interactive decision support tool to estimate priorities from pairwise comparison judgments. *International Transactions in Operational Research* 22, 2, 217–235.

van Til, J., Groothuis-Oudshoorn, C., Lieferink, M., Dolan, J., Goetghebeur, M., 2014. Does technique matter; a pilot study exploring weighting techniques for a multi-criteria decision support framework. *Cost Eff. Resour. Alloc.* 12, 1, 22.

Vanderbei, R.J., 2008. Linear Programming: Foundations and Extensions.

Vargas, L.G., 1990. An overview of the Analytic Hierarchy Process and its Applications. *Eur. J. Oper. Res.* 48, 1, 2–8.

Yeh, C.H., 2002. A problem-based selection of multi-attribute decision-making methods. *International Transactions in Operational Research* 9, 2, 169–181.