An Investigation of Outlying Performance in Menu Search

Hend Abdulrahman Albassam

Doctor of Philosophy

University of York

Computer Science

October 2021

Abstract

It has been noticed that in usability tests there is always one or more participants who take a long time to complete the task compared with other participants. Such participants appear as outliers in the collected time-on-task data. While such outliers can skew any statistical analysis, at the same time they may represent genuine problems with the interaction.

This research started with an exploratory study that investigated outlying performance in usability testing practices, to find out how outliers are interpreted and treated by practitioners. The practitioners interviewed in this study seem aware of the regular occurrence of outliers. They tend to link outlying performance cases to individual differences instead of linking them to usability problems. However, there appears to be no systematic approach to addressing them.

This research focuses on investigating outlying performance in menu search. Previous work suggests that the perceived menu semantics plays a role in outlying menu search performance. In this context, menu semantics refers both to the names of the menu items and to the names of the menu titles. Additionally, it refers to the semantic organization of the menu items. The series of studies conducted in this research used different methodologies to check whether the perceived menu semantics plays a role in outlying menu search performance. The results suggest that perceived menu semantics may play a role in outlying menu search performance.

Moreover, this research checked whether outlying menu search performance is due to specific individuals. The results suggest that outlying menu search performance can be due to individual differences.

The practical implication of this research is that outliers are a fact that should not be ignored. They should be considered and the reasons behind them should be identified. Consequently, based on the identified reasons, there is a possibility of improving the design for outliers.

Contents

List of Tables

List of Figures

Acknowledgements

In the name of Allah, the most gracious, the most merciful

First and foremost, I would like to express my thankfulness to Allah the Almighty for helping me to complete this thesis. Without His blessings, this work would never have been completed.

Special thanks go to my supervisor, Professor Paul Cairns. Thank you, Paul, for your guidance and encouragement throughout all stages of this research. I am most grateful for your patience and very valuable advice that inspired me to be positive. I have been really lucky to work under your supervision.

Above all, I am very grateful to my parents; they encouraged me and supported me in every phase of my PhD. I will never forget their unremitting prayers and their inquiries about my progress always pushed me to do more. My enthusiasm was instilled in me by my parents.

I would like to thank my husband, Waleed, from the bottom of my heart; he stood beside me on my long and tiring journey. Words cannot describe his role in and contribution to my PhD. Furthermore, I would like to thank my children, Sarah, Abdulilah and Basmah. It was not easy for them while I was very busy with my studies. I would also like to extend my deepest thanks to my parents-in-law, who constantly gave me enormous love and affection.

Last but not least, I cannot forget to thank my brothers, Ahmed and Abdullah, and my sisters Safaa, Hanaa, Fatmah and Razan for their help in my times of need. I am also grateful to my friend, Nouf Alrumaih, who was always there for me during the ups and downs of the PhD journey. In addition, great thanks must be extended to all those who participated in the studies. Without them, this research would never have been done.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

"6% of task attempts are extremely slow and constitute outliers in measured user performance. These sad incidents are caused by bad luck that designers can — and should — eradicate."

(Nielsen, 2006).

Chapter 1 Introduction

Usability tests are considered essential in the field of Human-Computer Interaction (HCI). Testing the usability of the systems is crucial. There have been many cases which show how the usability of a system has a huge impact on society as a whole including saving lives in a health context and preventing financial loss in a business context (Albert & Tullis, 2013).

Generally speaking, usability testing involves testing computer interfaces by using representative users to attempt representative tasks within a representative environment (Lewis, 2006). The goal of testing is to assess the usability of the system being developed by analysing how the intended users use the system to accomplish the tasks that the system was designed for (Preece et al., 2015). The usability evaluators use different metrics to measure usability, including performance metrics (Albert & Tullis, 2013). Performance metrics such as time-on-task are considered powerful tools when it comes to evaluating system effectiveness and efficiency. The time-on-task (the time needed to complete a task) is a commonly used performance metric in usability studies (Albert & Tullis, 2013; Hornbaek & Law, 2007). The time-on-task data are usually analysed by looking "…*at the average amount of time spent on any individual task or set of tasks by averaging all the times for each user by task*" (Albert & Tullis, 2013).

To compare the time-on-task data in different interface conditions, many usability studies use parametric statistical tests that assume that data is normally distributed such as t-tests and ANOVA (Cairns, 2007; Schiller & Cairns, 2008). It is common for researchers to use parametric tests even if the underlying assumptions such as normal distribution are not met. This is because they lack an understanding of these assumptions and do not use statistical tools that check them. In addition, they often believe that parametric tests are robust in terms of possible violations of the underlying assumptions (Mackenzie, 2013).

Cairns (2019) stated that assuming normality should not be a normal practice. In addition, he suggested that parametric tests should be abandoned as is not known when data are normal.

1.1 Is Performance Data Normal?

Normal data are typically linked to the classic bell curve distribution that has most data in the centre with decreasing amounts evenly distributed to the left and the right of the centre. In fact, real data cannot fit the bell curve as many collected data are not normally distributed. For example, time-on-task data is not normally distributed. The distribution of time-on-task data is typically positively skewed in usability studies (Albert & Tullis, 2013; Sauro & Lewis, 2010; Schiller & Cairns, 2008). The distribution of time-on-task data is expected to be skewed in this way because such data cannot be less than zero, but there are no upper limits. The positive skewness means that mean time will tend to be higher than median time and can, to some extent, be high enough to be not representative of typical user performance. This has direct effects when using parametric statistical tests which depend on the means of the samples.

However, because of the Central Limit Theorem, normal distributions dominate statistical methods that are used in HCI. This theorem states that regardless of the original population distribution, the sampling distribution of the means is almost normal if the sample is large enough. Even better, the mean of the sampling distribution of the means is equal to the population mean (Cairns, 2019). Therefore, parametric tests can be used without worrying about the normality of the population distribution as these tests just depend on the means. Based on the Central Limit Theorem, the means of time-on-task data can be normally distributed if the sample size is large enough. However, relying on the Central Limit Theorem can lead to some problems. The first problem is how large is a large enough sample. The theorem promises that the sampling distribution is normal in the end but does not tell where that will occur in terms of practical real data (Cairns, 2019; Wilcox, 2010). The second problem is that having large enough samples to provide a normal sampling distribution does not mean that the parametric tests behave as expected. For example, the t-test depends on an estimate of the standard deviation of the sampling distribution that is based on the standard deviation of the sample itself. If the sample is not normally distributed, the estimation of the standard deviation can be inaccurate, and therefore, the t-test can produce faulty results (Cairns, 2019).

Schiller and Cairns (2008) discussed the fact that the distribution of performance data (e.g., time-on-task) is not normal, but is positively skewed. Their discussion was based on an observation that in usability tests, there is always one or more participants who take a long time to complete the task compared with the other participants. Those participants appear as outliers in the collected time-on-task data. While such outliers can skew any statistical analysis, at the same time they may represent genuine problems with the interaction. In this research, an outlier is equivalent to the outlying performance of an individual.

1.2 Outlying Performance in Usability Tests

The term ‘outlier’ is commonly used in statistics. Howell (2016) defined an outlier as "*an extreme point that stands out from the rest of the distribution*". As statistics are commonly used in analysing the collected data in usability testing, the term ‘outliers’ is also used in this field. Albert et al. (2010) defined outliers in usability testing as *“extreme points in data and are the result of atypical behaviour”.* Barnum (2011) explained outliers in usability testing as follows *“an outlier is a single instance of a finding, something you have observed in one user only”.* The most common type of outlier in usability studies is time-on-task data when a participant takes a very long time to complete the task (Albert et al., 2010; Schiller and Cairns, 2008).

Outliers in HCI can be due to data entry errors, misbehaving participants, flawed design of the study and natural variability (Cairns, 2019). However, if outliers are not due to an external effect, be it errors, participant behaviour or study design, then outliers are valid data that should be analysed along with all other data (Cairns, 2019).

There are several methods for detecting outliers. The standard deviation (SD) method is one of the most frequently used methods for detecting outliers (Seo, 2006; Wilcox, 2010). The SD is a simple method that defines outliers as values that are more than two standard deviations from the mean. However, this method is sensitive to extreme values. The mean and the standard deviation can be inflated by the outliers and that can mask the presence of the outliers (Wilcox, 2010). Therefore, a more effective method for detecting outliers is required. The boxplot is a frequently used and commonly recommended method that is not sensitive to extreme values (Cairns, 2019; Wilcox, 2010). The boxplot uses quartiles to represent the distribution of the data. The lower quartile represents the value below which a quarter of the data exists, while the upper quartile represents the value that a quarter of the data is above. The difference between the upper and lower quartiles is the interquartile range (IQR) which shows the distribution of half of the data. The IQR for a normal distribution is approximately equivalent to $\frac{4}{3}$ standard deviations, therefore, is a good measure of spread. However, as it relies on quartiles, it is not modified by one or two outliers (Cairns, 2019). The common threshold used to identify outliers are values that are 1.5IQR over the upper quartile or under the lower quartile (Emerson and Strenio, 1983, as cited in Cairns, 2019). Both the SD and the boxplot are used in usability studies for detecting outliers. Cairns (2019) recommended using a boxplot as a method for detecting outliers as it offers a robust method to identify outliers in data representing a single variable (Cairns, 2019).

Outliers should be considered for two reasons. Firstly, outliers are considered a serious problem for statistical analysis (McClelland, 2000). The regular occurrences of outliers in the data lead to the distribution does not fit with a normal distribution. Consequently, the parametric statistical tests such as t-tests that are commonly used in usability studies, are directly influenced because they assume the data is normally distributed. Thus, the results of these tests are potentially not valid (Cairns, 2019). Nonparametric statistical tests can be used when the data is not normally distributed (Lazar et al., 2017). These nonparametric tests are robust to outliers because they depend on ranking the data, while parametric tests rely on the means of the samples, which can be modified due to the outliers (Cairns, 2019). Nonparametric statistics might help in solving some problems of analysis, but they cannot solve the fact that outliers might represent a genuine problem in the interaction.

Secondly, outliers can provide useful information about the collected data. Outliers might indicate usability problems. For example, Law and Hvannberg (2004) and Nielsen and Landauer (1993) noticed that in usability testing, a significant number of usability problems were identified by just one user. In addition, outliers might indicate individual differences, as was found by Egan (1988), who suggested that individual differences in HCI can lead to significant variance in user performance. These individual differences should be considered as they can be related to some HCI problems (Dillon & Watson, 1996).

Unfortunately, outliers seemed to be overlooked in usability studies, and the focus of any analysis is typically more on average users who are more representative of the population as a whole. There is a tendency to disregard outlying performance in usability tests as it tends to be observed in just one participant (Barnum, 2011). Ignoring outliers and not taking them seriously can lead to both statistical problems and the possibility of overlooking usability problems.

Apparently, outliers are a potentially important feature of usability tests and not just a statistical nuisance. Their regular occurrence and their potentially negative consequences in usability studies necessitate thinking about them rationally. However, a lack of research investigating such a situation makes it hard to know the extent or impact of such a problem on usability test results. Therefore, there is a need to explore the causes of outlying performance and based on the identified causes, there is a chance to improve the design for outliers.

1.3 Outlying Performance in Menu Search

Schiller and Cairns (2008) identified outliers in menu search and, based on modelling work, attributed this to the perceived menu semantics. They proposed a cognitive model that simulated a user searching a menu hierarchy. Their proposed model used the information scent of the relevance of a menu item to the given task to decide on actions to perform. For example, in a perfectly designed menu, the user considers only the target item to be relevant and therefore scented with a value of 5, while all other items are irrelevant and therefore scented with a value of 1. Their cognitive model rejects all the 1-scented items and selects the 5-scented item. Schiller and Cairns suggested that people may vary in terms of perceiving the menu semantics, and these variations lead to different menu search behaviours. By reflecting the variations in perceiving the menu semantics in their proposed menu search model, the model was able to predict outliers as a result of these variations.

Menu semantics is a key factor in menu search performance. It refers to the names of the items. Also, it refers to the names of the menu titles. Additionally, it refers to the semantic organization of menu items (Baily et al., 2016). The way in which menu items are grouped and the titles used to name the groups critically influence the menu search performance (Lee & Raymond, 1993). It is commonly assumed that a menu is efficient when its items are organised in such a way to match the users' perception of menu semantics (Schmettow & Sommer, 2016). Therefore, menu designers used a number of methods such as card sorting to elicit users' perception of menu semantics. Users might be different when it comes to perceiving menu semantics and this might justify outlying performance in menu search as suggested by Schiller and Cairns (2008).

If outlying performance is due to the perceived menu semantics, outlying performance might be fixed in this case. Fixing outlying performance through improvements in the menu design does not necessarily improve the average user performance. However, the performance of outliers could be improved significantly by improving the menu design. Therefore, there is an opportunity to help outliers and improve the quality of their interaction.

The aforementioned work of Schiller and Cairns (2008) motivated this research to investigate outlying performance in the interaction with menus. Menus are a common interaction method used in current systems by various types of users. They are used in many applications and systems to allow the user to navigate and select the target item in a structured way (Bailly et al., 2016). They exist in desktop applications, on websites, smartphones and tablets. In addition, they are used in home control systems and medical devices. Regardless of their uses, menus should offer the user a quick and easy way to find and select a target item (Brumby & Zhaung, 2015). Therefore, it is important to investigate the possible causes of outlying performance in menu search.

This research aims to investigate outlying menu search performance. In this context, outlying menu search performance refers to one or more participants who take a long time to complete an individual menu search task compared with other participants. As a result, such participants appear as outliers in the boxplot of the distribution of the collected menu search time data.

1.4 Research Questions

To motivate investigating outlying performance in general, it is important to check whether usability practitioners take outliers seriously and whether they have ways of dealing with them. Moreover, it is important to know how practitioners interpret outliers in usability tests as this might help in framing the findings of the studies that investigate outlying performance empirically in this research. Therefore, this research helps to answer the following research question:

(RQ1) How are outliers interpreted and treated in usability testing practice?

As outlined before, this research focused on investigating outlying menu search performance. Schiller and Cairns (2008) attributed outlying menu search performance to the perceived menu semantics based on modelling work. However, modelling alone is not an alternative for gathering and analysing data on real users engaged in searching menus, which provides empirical evidence that backs any claim. Therefore, there is a need to address this limitation by conducting empirical studies that check whether the perceived menu semantics plays a role in outlying performance in menu search. Therefore, this research addresses the following research question:

(RQ2) Does the perceived menu semantics play a role in outlying performance in menu search?

Additionally, outlying menu search performance might be due to individual differences in cognitive and psychomotor abilities, and traits. Individual differences in cognitive and psychomotor abilities (e.g., reaction speed and hand dexterity) might cause variability in performance with regard to using computers (Kuurstra, 2015). In addition, individual differences in traits (e.g., the Big Five personality dimensions: extraversion, neuroticism, openness to experience, conscientiousness, and agreeableness) can produce variability in performance with regard to using computers (Kuurstra, 2015). If outlying performance is due to a permanent feature in a participant, the same participant will show outlying performance on regular basis. Therefore, there is a need to check whether outliers occur on a regular basis. Therefore, this research helps to answer the following research question:

(RQ3) Does outlying performance in menu search occur due to specific individuals?

1.5 Research Approach, Scope, and Methods

This research aimed to investigate outlying performance in menu search. It was mainly based on a series of empirical studies to address the research problem. These studies have employed different methodologies to collect and analyse the data. The methodologies employed included qualitative and quantitative methods. Semi-structured interviews and retrospective think-aloud (RTA) protocol methods were used to collect qualitative data. The thematic analysis method was used to analyse the collected qualitative data. Controlled experiments, surveys, modelling and card sorting methods were used to collect the quantitative data. Statistical analysis methods were used to analyse the collected quantitative data.

This research started with an exploratory study (Study 1- Chapter 3) to answer the first research question that asks how outliers are interpreted and treated in usability testing practice. This study involved interviewing practitioners to collect data related to how they interpret and treat outlying user performance cases in usability testing.

As this research focuses on investigating outlying performance in menu search, Study 2 to Study 7 investigated outlying performance in menu search using different methods. The menu layout used in these studies was a linear one where items are organised vertically. This layout was used because the majority of menus are displayed linearly (Bailly et al., 2016). The menu search task in these studies involved asking participants to find and select a specific target from the menu as quickly and accurately as possible. A boxplot was used in these studies to visualize the distribution of the collected menu search time data and to identify outliers.

To answer the second research question that asks about the role of the perceived menu semantics in outlying performance in menu search. Study 2 (Chapter 4) was conducted to investigate whether poor semantic organization of menu items plays a role in outlying performance. A between-group design was used in this study; two conditions of menu semantic organisation were identified for testing in this study: a semantically organised menu and a randomly organised menu. The participants were randomly assigned to two groups and asked to select a specific target item from the displayed menu.

To further check the role of the perceived menu semantics in outlying performance, a menu search model was adopted to help in understanding the role of the perceived menu semantics. This model is based on machine learning (ML) and needs to be trained on menu samples to learn menu search strategies. Constructing menu samples needs the collection of semantic similarity ratings from human participants. Therefore, Studies 3 and 4 (Chapter 5) were conducted to collect the semantic similarity ratings of menu items. Study 3 used the pairwise similarity ratings method to collect the semantic similarity ratings. Study 4 used a card sorting method to collect the semantic similarity ratings. Two methods were used to collect the semantic similarity ratings to see whether they introduced different features on the data. In Study 3, the participants were given a survey that consisted of a list of pairs of menu items. They were then asked to rate each pair according to what they thought about how close in meaning the two menu items in the pair were. In Study 4, the participants were asked to sort the menu items into groups of similar items. After that, Study 5 (Chapter 6) was conducted to train and test the adopted menu search model using menu samples constructed from the collected semantic similarity ratings.

Following this, Study 6 (Chapter 7) was conducted to investigate whether participants who have a different perception of menu semantics performed poorly in the menu search task. In addition, this study tried to answer the third research question by investigating whether outlying performance is due to specific individuals. The participants were asked to perform two tasks: a card sorting task and a menu search task. In the card sorting task, the participants' menu semantics perception was elicited. In the menu search task, the participants were given several menu search trials, and their menu search performance in each trial was measured. Outliers were found in each menu search trial. However, the reasons behind their poor performance were not clear.

It was difficult to determine the reasons for outlying performance cases in Study 6. Therefore, a think-aloud protocol (TA) was suggested to help determine the reasons. Study 7 (Chapter 9) was conducted to investigate outlying performance in menu search using a retrospective think-aloud protocol (RTA). The participants were asked to perform menu search tasks. They were then shown a screen recording of their performance in these tasks and were asked to verbalise their thoughts while undertaking these tasks.

1.6 Research Contributions

The focus of this research, as explained earlier, was on investigating outlying performance in menu search. The series of empirical studies reported in this thesis contributes to our general understanding of outlying performance and why it might happen in menu search.

The key findings of this research are:

- The interviewed usability practitioners seem aware of the regular occurrence of outliers in usability testing. They tend to link outlying performance cases to individual differences instead of linking them to usability problems. However, there appears to be no systematic approach to addressing them.

- There is always one or more participants (outliers) who take a long time to complete the menu search task compared with other participants.
- The results suggest that the perceived menu semantics may play a role in outlying menu search performance.
- The results suggest that outlying menu search performance can be due to individual differences.

1.7 Ethical Statement

All the studies conducted in this research followed the principles of ethical research involving humans. All the studies reported in this thesis adhered to the ethical principles of ‘Do No Harm’, ‘Confidentiality’, and ‘Informed Consent’. All the conducted studies were approved by the Physical Sciences Ethics Committee (PSEC) of the University of York and followed the ethics procedures of the Department of Computer Science.

Do No Harm

No one participant in any of the studies conducted in this research was exposed to any harmful environments. Participants were free to withdraw from the study at any time. They were told that they would not be affected negatively by withdrawing, and they would not be asked about the reasons behind their withdrawal.

Additionally, this work is about exploring the causes of outlying menu search performance. This might help in improving the menu design for all users. Therefore, the research findings are not intrinsically likely to have a harmful impact if they are applied in the real world.

Confidentiality

The anonymity and confidentiality of the collected data in all studies were maintained. The anonymity of participants was always maintained when reporting the results, and each participant was given an ID to be used when quoting participants’ words. The collected data were kept in password-protected systems that prevent unauthorised access to these data.

Informed Consent

The participants in all studies were given a brief information about the study that they had been invited to take part in. In Studies 1,2,3,4, and 7, which were conducted face-to-face, an information sheet was given to the participants to inform them about the aims and procedures, and their rights, before the study began. In addition, an informed consent form was given to the participants to obtain their consent to participate in the study before the study started. After completing the study, the participants were debriefed. Study 6 was conducted online. The participants in this study were given brief information about the study and their rights before the study started, and were allowed to contact the researcher by email if they had any inquiries.

Chapter 2 Literature Review

2.1 Introduction

This thesis is motivated by the observation that there are always one or more outlying user performance cases in usability tests. Therefore, this literature review begins by providing an overview of usability testing and measuring user performance in usability testing. It presents performance metrics and focuses particularly on a time-on-task metric. Then, the review moves to outlying user performance in usability testing. Therefore, it defines outliers and explains how outliers are detected in statistics. It also discusses how outliers are interpreted and treated in usability studies. Additionally, it reviews the relevant works that investigated outlying user performance.

This thesis focuses on outlying performance in menu search. Therefore, the review defines menus and presents menu properties. Menu semantics is an influential factor in menu search performance, the perceived menu semantics was suggested by previous work as a factor in outlying performance in menu search. Therefore, the review presented some studies on the effect of menu semantics on user performance in menu search.

Additionally, the review delivers an overview of the existing menu search models and provides several details about the adopted model in this thesis to model outlying menu search performance. Furthermore, it explains the methods used to collect the semantic similarity ratings of menu items that are needed to run the adopted menu search model.

Finally, the literature review concludes by discussing individual differences and user performance in usability studies as a possible cause for outlying performance in usability testing.

2.2 Usability Testing

The International Standard Organization (ISO 9241-11) defines usability as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use*." (ISO, 2018). According to Albert and Tullis (2013), there is a difference between usability and user experience, "*Usability is usually considered the ability of the user to use the thing to carry out a task successfully, whereas user experience takes a broader view, looking* *at the individual's entire interaction with the thing, as well as the thoughts, feelings, and perceptions that result from that interaction".*

Usability testing is frequently regarded as the gold standard of usability evaluation methods (UEMs) (Hornbæk, 2010). Usability testing involves testing computer interfaces by bringing in typical users and asking them to attempt typical tasks within typical environments (Lewis, 2006). It plays a vital role in improving the quality of an interface by identifying the weaknesses in an interface design that cause problems for users. Additionally, it can be used as a user research method to understand how users interact with interfaces (Lazar et al., 2017).

Usability testing can be divided into two types based on the goal of usability of the study and when it is done: formative testing and summative testing. Formative testing is conducted during the development of the product and aims at identifying and fixing problems. It usually consists of small repetitive studies. Summative testing is conducted after finishing the product and aims at creating a baseline of metrics or confirming that the requirements are met by the product. It typically needs a large number to provide statistically validated results (Barnum, 2011).

Usability testing can be conducted anywhere. It can be conducted in a fixed lab, an office, a user's house, via the phone, or via the web. None of these locations is better than the others. The decision should be based on the type of usability testing project (Lazar et al., 2017).

During usability testing, usability evaluators use usability metrics to measure usability. These metrics should be observable by some means to assess the user's interaction with a system and reveal *"...some aspect of effectiveness (being able to complete a task), efficiency (the amount of effort required to complete the task), or satisfaction (the degree to which the user was happy with his or her experience while performing the task)"* (Albert & Tullis, 2013). Additionally, usability metrics need to be quantifiable, so they can be converted to a number and counted somehow. All metrics need the issue being evaluated to be stated numerically. For example, a usability metric can show that 90% of the users completed the tasks in less than 1 minute (Albert & Tullis, 2013).

Usability metrics can be divided into several categories including performance metrics, issues-based metrics, self-reported metrics, behavioural and psychological metrics, and combined and comparative metrics (Albert & Tullis, 2013). As the name suggests, performance metrics are the most suitable for investigating outlying user performance, and therefore, they are covered in more detail in the next section.

2.3 Measuring User Performance

Albert and Tullis (2013) stated that during the usability testing, the user's performance is considered an important indicator of general usability. Therefore, performance metrics are considered robust tools to evaluate the effectiveness and efficiency of different systems. For example, by identifying users who are making many errors, it is hypothesised that there is potential for improvement. Moreover, by finding the users that spend a long time completing a task compared with what was estimated, it is realised that efficiency can be improved significantly.

Despite the advantages of performance metrics, some issues should be considered when collecting performance metrics. First, the sample size should be sufficient; 10 participants or more are required to obtain meaningful results with reasonable confidence levels. Also, performance metrics cannot be used alone to discover usability problems. The collected performance data can help in pointing to tasks or elements of an interface that were problematic. However, they cannot help in identifying the sources of the problems. Therefore, performance data is commonly supplemented with other data such as observational or self-reported data to gain a better understanding of the reasons behind the problem and suggest solutions to fix them (Albert & Tullis, 2013).

There are five basic types of performance metrics (Albert & Tullis, 2013):

- Task success measures the percentage of users who completed a task successfully.
- Time-on-task measures the time that the user needed to complete a task.
- Errors measure the number of mistakes that the user made during a task.
- Efficiency measures the effort that the user spent to complete a task.
- Learnability measures improvement in user performance over time.

Time-on-task (sometimes called task completion time or task time) is a commonly used performance metric in usability studies (Albert & Tullis, 2013; Hornbæk & Law, 2007; Sauro & Lewis, 2009). This is because it is a good approach to evaluate the efficiency of a system (Albert & Tullis, 2013). Also, it is a good way to discover usability problems as a long time on task might indicate problems in the interaction (Sauro, 2011).

Previous analysis of usability studies found that faster task completion time was correlated with fewer errors and greater satisfaction (Sauro & Lewis, 2009). In most cases, a faster task completion time indicates a better experience. However, there are some exceptions to this assumption. For example, games where users want to spend more time to enjoy as well as e-learning systems where users gain more if they spend more time (Albert & Tullis, 2013).

Time-on-task is the time passed between starting a task and finishing a task. It is expressed in minutes or seconds. It can be measured in several manual or automated ways. In the manual methods, the moderator simply records the start and end times; also, he can use a stopwatch or any time-keeping tool to measure the time. However, there are many automated tools that allow recording the time more accurately and less obtrusively (Albert & Tullis, 2013).

One of the issues that should be considered during measuring time-on-task is whether to inform the participants about the time measurement. Participants likely feel nervous if they know that their time is being recorded, and this affects their performance. On the other hand, if they are not told, they might take their time exploring the interface. A good balance is to ask the participants to complete the task as quickly and accurately as possible without mentioning that time is recorded (Albert & Tullis, 2013).

There are many ways to present and analyse the collected time-on-task data. The most common way is to look at the average amount of time spent on any individual task or set of tasks by averaging all the times for each user by task (Albert & Tullis, 2013). This is a simple and intuitive way to analyse time-on-task data. The distribution of time-on-task data is typically positively skewed in usability studies. Therefore, the median is reported by practitioners who are aware of this skewness (Sauro & Lewis, 2010).

Schiller and Cairns (2008) discussed the fact that the distribution of user performance data (e.g., time-on-task) is not normal but positively skewed. Their discussion was based on an observation that in usability tests, there is always one or more participants who take a long time to complete the task compared with other participants. Those participants appear as outliers in the collected time-on-task data. Outliers can skew any statistical analysis but, at the same time, may represent genuine problems with the interaction. Because of this, outlying user performance is potentially an important feature of usability tests, not just a statistical nuisance.

2.4 Outlying Performance in Usability Tests

To investigate outlying performance, there is a need to know the definition of outliers and how they are interpreted and treated in usability studies. Additionally, the relevant works, that investigated outlying performance in usability studies, need to be reviewed.

2.4.1 Definition of Outliers

Aguinis et al. (2013) collected 14 definitions of outliers that were used in the methodological sources. These definitions are different, in some definitions, outliers are extreme values that are remarkably distant from the central tendency, while in others, in addition to being distant from the central tendency, outliers must either disturb the results or produce some useful or unforeseen insights. Additionally, in some definitions, outliers are not depending on any type of data analysis, while in others, outliers are values that disturb the results of a particular type of data analysis (Leys et al., 2019).

Leys et al. (2019) suggested that two of these 14 definitions seemed very appropriate for practical reasons. The first one is appealing for its simplicity: *" Data values that are unusually large or small compared to the other values of the same construct"*. However, this definition can be just used with a single construct, multivariate outliers, that are caused by an unexpected pattern across multiple variables, should be also considered. Therefore, another comprehensive definition can be relied on: *"Data points with large residual values"*.

The term 'outliers' is used in the usability testing field. Albert et al. (2010) defined outliers in usability testing as *"extreme points in data and are the result of atypical behaviour"*. Barnum (2011) explained an outlier in usability testing as follows *"An outlier is a single instance of a finding, something you have observed in one user only"*.

2.4.2 Outliers Detection Methods

Howell’s definition of an outlier stated that an outlier is an extreme point that is far away from the rest of the data. However, where is the rest of the data? And what is far away? Therefore, there is a need to work out where the rest of the data is to define where outliers are not. A measure of central tendency and a measure of spread such as the usual mean and the standard deviation can help. Therefore, in statistics, typically, the default method for defining outliers is in terms of the normal distribution. The mean represents the average performance, and the standard deviation represents the amount of variation (Cairns, 2019).

The Standard Deviation (SD) method uses the mean and the standard deviation to define outliers. It is one of the most frequently used methods to detect outliers (Seo, 2006; Wilcox, 2010). It is a simple method that defines outliers as values that are more than two standard deviations from the mean. However, this method is sensitive to extreme values. The mean and the standard deviation can be inflated by the outliers and that can mask the presence of the outliers (Wilcox, 2010).

Another method that uses the mean and the standard deviation to detect outliers is the *z*-score. By transforming the data to *z*-scores, which measure the number of standard deviations below or above the mean, a data point is, it can be checked if a data point is far away from others. To declare that the data point is outlying, a threshold is set, and this is decided based on what would be thought rare as for the normal distribution. For example, Tabachnik and Fidell (2001, as cited in Cairns, 2019) recommend a threshold of $z > 3.29$, matching a p-value $p < 0.001$. The *Z*-score method is widely used to define outliers in psychology (Bakker and Wicherts, 2014). However, Cairns (2019) argued against the use of z-scores as a statistical method to define outliers and stated it is flawed. This is because both mean and standard deviation are calculated using outlying data points, and an outlier causes moving the mean headed for itself and increases the standard deviation. These effects lead to reducing the z-score and consequently reducing the seeming severity of an outlier.

A better and recommended method that is not sensitive to extreme values is a boxplot (Cairns, 2019; Wilcox, 2010). The boxplot uses quartiles to represent the distribution of the data, see *Figure* 2.1. The lower quartile represents the value that a quarter of the data is under this value and the upper quartile represents the value that a quarter of the data is above this value. The difference between the upper and lower quartiles is the interquartile range (IQR) which shows the distribution of half of the data. The IQR for a normal distribution is approximately equivalent to $\frac{4}{3}$ standard deviations, therefore, it is a good measure of spread, but as it relies on quartiles, it is not modified by one or two outliers (Cairns, 2019). The common threshold to identify outliers are values that are 1.5IQR over the upper quartile or under the lower quartile (Emerson and Strenio, 1983, as cited in Cairns, 2019). This matches the 2.7 z-score in a normal distribution (Cairns, 2019). The outliers are shown in a boxplot as separate circles, see Figure 2.1.

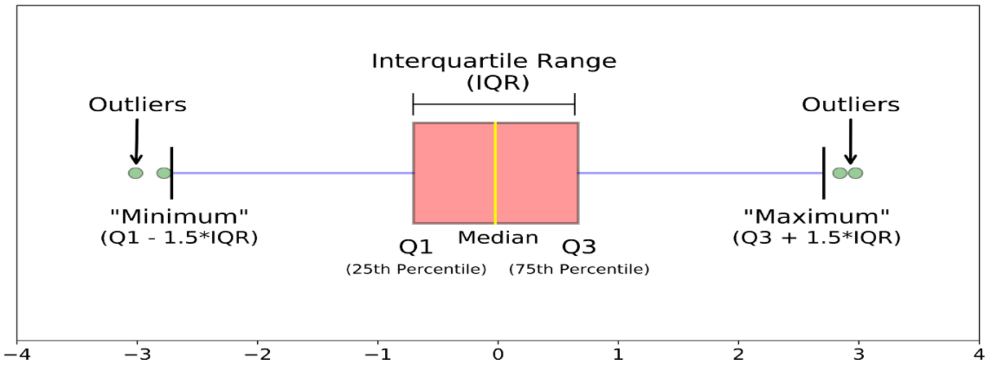


Figure 2.1: Identifying outliers by using the boxplot (Source: Galarnyk, 2018).

Another method to robustly detect outliers is called Median Absolute Deviation (MAD) (Cairns, 2019). It is calculated by subtracting the median from each observation and then taking the absolute values. MAD is the median of these calculated values (Wilcox, 2010). Outliers are defined as points that are a certain number of MADs from the median. This threshold is usually set at 3 (Cairns, 2019).

According to Cairns (2019), the MAD is similar to the IQR in that 50% of the data is within 1 MAD of the median, but the MAD is symmetrical about the median by definition, whereas the box on a boxplot can be asymmetric. When the underlying distribution is known to be asymmetric around the median, the MAD produces a lower threshold than the IQR, allowing more data to be defined as outlying. Therefore, Cairns (2019) prefers the IQR way to define outliers, because it shows that distributions can be asymmetric and so outliers can be also asymmetrically distributed.

All the aforementioned methods for detecting outliers are suitable when the data distribution is symmetrical, and bell-shaped such as with the normal distribution (Seo, 2006). The boxplot and the MAD may be appropriate for skewed data (Seo, 2006), although, for very large datasets, other methods may be more appropriate.

In the HCI context, the SD method is used in usability studies as a standard way to detect outliers (Albert et al., 2010; Albert & Tullis, 2013). Also, the boxplot method is used to detect outliers in the collected data in usability studies such as Glasser (2019), Mtonga et al. (2018), and Robertson and Kortum (2020).

The boxplot is a well-established method to represent data, including outlying data points (Cairns, 2019). It is used in usability studies to represent the collected time-on-task data such as Mayer et al. (2012), Iftikhar et al. (2021), and Ehrler et al. (2018). Therefore, all studies in this research that collected the time-on-task data, used the boxplot to visualize the distribution of the collected time-on-task data and to identify outliers.

2.4.3 Interpreting and Treating Outliers in Usability Studies

Outliers seemed to be overlooked in the usability studies, and the focus of any analysis is typically more on average users who represent the population. There is a tendency to disregard outlying performance in usability tests as it is just observed in one participant (Barnum, 2011).

However, recent work by Cairns (2019) discussed the actual causes of outliers in HCI and hence what should be done to deal with them. He recommended that when having outliers, several basic causes of outliers should be checked to determine the reasons behind those outliers. The basic causes of outliers in HCI include erroneous data, misbehaving participants, flawed design of the study and natural variability.

Erroneous data is one of the basic causes of outliers in HCI. Outliers might be caused by errors during data entry, in this case, there is a need to check everything to make sure that there are no other errors in data entry (Cairns, 2019).

Misbehaving participants is also considered a basic cause of outlying performance in usability tests. Outliers might suggest that participants did not behave appropriately in the study, therefore, there is a need to identify and remove all similar misbehaving participants (Cairns, 2019).

Additionally, the faulty study design is one of the basic causes of outliers in HCI. Outliers might indicate a flaw in the study design including recruiting a participant who does not fit with the rest of the participants, in this case, all similar not fitting participants should be identified and excluded from the study (Cairns, 2019).

When outliers are found due to the abovementioned external effects, all data should be checked against this influence. However, if none of the abovementioned external effects is found, then outliers are valid data that should be analysed along with all other data (Cairns, 2019).

Folstad et al. (2012) discussed treating usability problems found for only one participant. These are called single-user problems. They stated that deciding on single-user problems is challenging due to the difficulty to assess their relevance and validity. Moreover, the literature does not help in how to tackle this issue and how to decide to accept or reject these problems as usability problems. They discussed the conflicting opinions on how to handle single-user problems in usability testing. For example, Dumas and Redish (1993) suggested reporting single-user problems as outliers. Kjeldskov et al. (2004) argued for considering single problems as noise instead of usability problems. In contrast, Woolrych and Cockton (2001) considered the analysis of a "stress test", a usability test designed to specifically address problems identified through heuristic evaluation. They found that, depending on which 5 participants might have taken part, single-user problems could still indicate severe usability problems. Folstad et al. (2012) stated that no study discussed single-user problems in detail to advise the practitioners on relevant aspects such as consideration of the size of the sample, the estimation of problem re-occurrence and the likely test-specific factors. Therefore, they investigated treating single-user problems in practical usability testing. They conducted an online survey. They asked the usability practitioners to respond to the question: "*If an incident was observed with only one of the users participating in your latest usability test, how did you decide whether this was a usability problem or not?*". Eighty-nine (89) usability practitioners working in 17 countries responded to the survey. It was found that usability practitioners varied in treating these single-user problems. And based on usability practitioners' responses, Folstad et al. (2012) proposed five recommendations on how to treat a single user problem. First, setting a procedure to deal with single-user problems. Second, considering the sample size as a single user problem in a small sample size should be given more attention. Third, using knowledge resources including heuristics, guidelines, and results from previous tests to evaluate a single user problem. Fourth, asking for advice from experts and other group members. Fifth, it is not necessary that a single user problem indicates usability problems; it might be justified as an artefact of test settings. This multi-faceted response supports Woolrych and Cockton's view that a single method of usability testing is not sufficient.

Outliers might be considered a specific type of single-user problem: one user who takes a particularly long time, for instance. However, single-user problems represent distinctions between the problems encountered by individuals, whereas outliers are people who perform differently on the same tasks and measures as everyone else. Thus, it may be due to them encountering specific distinct problems that other users did not encounter or may be due to other causes. As such, they may need similarly careful, distinct consideration when interpreting the outcomes of a usability test.

2.4.4 Investigations of Outlying User Performance

There has been very little empirical research investigating outlying user performance in usability studies; only three relevant studies were found and reviewed.

The first study was conducted by Schiller and Cairns (2008). They discussed why user performance, specifically time-on-task, is not accurately represented by a normal distribution. This is because informal analysis of many usability studies showed that there is always one participant or more who performed extremely poorly compared to others. This participant appears as an outlier in the collected data. The regular occurrences of outliers in the data lead to the distribution does not fit with a normal distribution. To find a better statistical distribution of user performance data, they took a modelling approach. They implemented a cognitive model called TreeWalker. This model is based on the cognitive model of Cox and Young (2004) that models user behaviour during menu exploration to select an item from this menu. TreeWalker model extended Cox and Young's model and simulated a user searching a more complicated menu hierarchy. In the first part of their study, they recruited many participants online to collect relevance ratings of menu items to the task based on a scale of 5 (very relevant) to 1 (not at all relevant). Also, they asked the participants to complete menu navigation tasks using menus from a real-world website, and their task completion time was recorded. After that, in the second part of the study, the model was parameterised with relevance rating data collected from the participants. Then, the model was run, and the results were compared with the recorded task completion time. The distribution of performance data and the percentage of outliers generated by the model fit well with the data obtained from the participants in the first part of this study. Two methods were used to identify the outliers. The first method used a standard estimate of mean and standard deviation. The second method used a robust method based on medians and deviations from the median. Schiller and Cairns' study and their proposed model showed that outlying performance was occurring more frequently compared with a normal distribution expectation. Outliers that are generated by the theory should not be removed; they should be examined carefully as they might indicate an important insight. Schiller and Cairns found that their model predicted outlying performance as a result of the perceived menu semantics.

The second study was conducted by Auskerin (2012) to build and extend the evidence of persistent outliers that was highlighted by Schiller and Cairns (2008). A following sensible step was decided to carry out an actual usability test to collect cases of outlying user performance. Therefore, a prototype of an e-commerce website was designed and built, and an experiment consisting of several menu navigation tasks was conducted. Performance metrics including task completion time, page views and efficiency were used to measure user performance and identify outliers accordingly. Many cases of outlying user performance were observed and confirmed Schiller and Cairns' observations. However, the underlying causes of the outlying performance could not be discovered, and based on the initial analysis, age and computer experience were suggested as possible causes of outlying performance.

The third work was done by Yin[1] (2018). She investigated whether individuals who rated the semantic similarity of menu items differently performed poorly in menu search tasks. She also investigated whether an outlying performance was due to specific people. She conducted two experiments in her investigation. In both experiments, two tasks were to be carried out by the participants: the similarity rating task and the menu navigation task. Both experiments used the pairwise similarity ratings method to collect the semantic similarity ratings of menu items. The difference

[1] Yin built her work on this research. She built her work based on Study 2 (Chapter 4) and Study 3 (Chapter 5) in this research that showed a clue for a possible reason for outlying performance in menu search. This reason was individuals who rated the semantic similarity of menu items differently.

between these experiments was in the menu navigation task. In the first experiment, the participants were asked to do a single search task in a single level menu, while in the second experiment, the participants were asked to perform multiple menu search tasks in a two-level menu. multiple menu search tasks were designed to check whether outliers occur on a regular basis. She found neither similarity ratings nor the person responsible for the outlying performance. The occurred outliers in the menu search task were not different from others in the semantic similarity ratings of menu items. Additionally, the occurred outliers were not due to specific people as they did not show outlying performance on a regular basis.

The aforementioned studies provided preliminary results that motivated this research to go further and investigate the outlying performance in menu search. For example, Schiller and Cairns' work identified outliers in the menu search and attributed this to the perceived menu semantics as a result of some modelling work. However, modelling alone is not an alternative for gathering and analysing data on real users searching menus, which provides empirical evidence that backs any claim. Therefore, there is a need to address that limitation by conducting empirical studies that check whether the perceived menu semantics plays a role in outlying menu search performance.

As this research is about outlying user performance in menu search, the focus in the next section is on user performance in the interaction with menus.

2.5 User Performance in Menu Search Tasks

Menus exist in many applications and systems to allow the user to navigate and select the target item in a structured way. They exist in desktop applications, on websites, smartphones and tablets. Also, they are used in home control systems and medical devices. Regardless of their uses, menus should offer the users a quicker and easier way to find and select the target item (Brumby & Zhaung, 2015). Many design aspects affect user performance during interacting with menus. Therefore, it is important to understand these aspects and their effect on user performance in menu search tasks.

2.5.1 Definition of Menus

According to Baily et al. (2016), while the terms "menu", "menu system", and "menu technique" are commonly used, there is no agreement on the definition of these terms in the literature. There have been many suggested definitions. For example, ISO (1998) defines a menu as "*a set of selectable options*". Norman (1991) defines menu techniques as "*Menu selection is a mechanism for users to indicate their choices. The characteristics of menu selection are that a) the interaction is, in part, guided by the computer; b) the user does not have to recall commands from memory, and c) user response input is generally straightforward".* Helander et al. (1997) define menus as *"a set of options, displayed on the screen, where the selection and execution of one (or more) of the options result in a change in the state of the interface".* Baily et al. (2016) stated that these definitions are general to some extent, and it is challenging to define a menu precisely. Therefore, they proposed four key characteristics and considered them for defining menus:

- Menus enable users to select commands from a set of items (ISO, 1991; Foley, 1999).
- Menus present items in a structured visualisation. Items are generally organised in hierarchical groups or categories. These groups or categories might be demarcated by separators. Items might be arranged alphabetically, numerically, semantically or based on their frequency of use. Items are positioned based on a geometrical structure (linear, circular, etc.) that allows users to find target items (Dachselt, 2007; Jackoby& Ellis, 1992).
- Menus are transient (Jackobsen et al., 2007). Transient visualisations allow the information to be temporarily presented and easily terminated. Menus do not need persistent screen space as they show up when needed and closed directly after selecting an item.
- Menus are quasimodal (Raskin, 2000). Quasimodes are modes *"that are kept in place only through some constant action on the part of the user"* (Raskin 2000). Once a menu is activated by the user, the system gets into a specific mode till completing the selection task.

Additionally, Baily et al. (2016) presented a taxonomy of menu properties based on three dimensions: item, menu and menu system. The dimensions are arranged hierarchically: items belong to menus that belong to a menu system, see Table 2.1.

The same property can mean a different thing for each dimension. For example, the geometry property of an item is related to its size and position, while for a menu, it is related to its layout as well as the layout of items inside it. For semantics property, it is also related to different things for each dimension. The semantics of an item is about its name, while the semantics of a menu is related to the wording within a menu and the menu's title. For a menu system, semantics property is about the semantic organisation of menu items with the corresponding menu hierarchies (Baily et al., 2016).

Dimension	**Subdimension**
Item	Geometry
	Visual representation
	Semantics
Menu	Geometry
	Temporality
	Semantics
Menu System	Semantics
	Menu depth
	Menu breadth

Table 2.1: The Taxonomy organises menu properties based on three dimensions (Source: Baily et al., 2016).

Menu semantics is a key factor in menu search performance. As mentioned before, menu semantics refers to the names of the items, so the name of a menu item should be relevant. Also, menu semantics refers to the names of the menu titles, so the title should reflect the items in the corresponding submenu. Additionally, menu semantics refers to the semantic organization of menu items, so the menu items should be organized into meaningful hierarchies (Baily et al., 2016).

Brumby and Zhaung (2015) provided examples of semantically organized menus. These examples are the Settings menu in the Apple iPad and the File menu in Microsoft Word, see Figure 2.2. These menus put functionally related items in one group, for instance, Airplane Mode, WiFi, and Bluetooth are related to managing radios, while the next set of items is related to managing notifications.

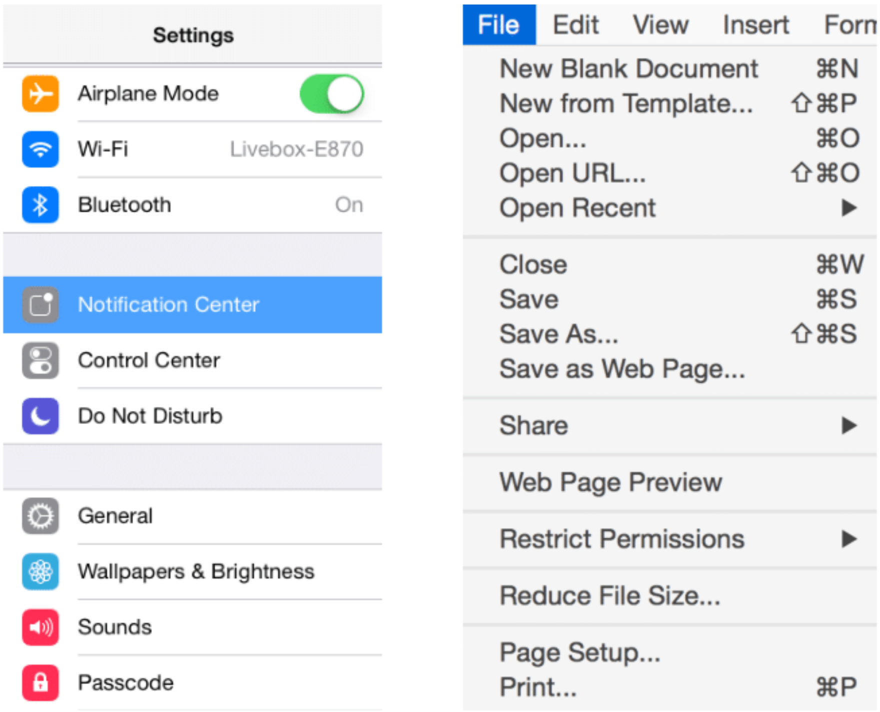


Figure 2.2: Examples of semantically organized menus, (a) the Settings menu in the Apple iPad, (b) The File menu in the Microsoft Word (Source: Brumby and Zhaung, 2015).

The effect of menu semantics has been studied in several studies. Therefore, in the next section, the experimental studies that focus on menu semantics were reviewed to further understand the relation between menu semantics and user performance.

2.5.2 The Effect of Menu Semantics on Menu Search Performance

The effect of the semantic organisation of menu items on user performance during searching menus was explored by an early work by Card (1982). He conducted a study to understand how users search the computer command menu. The study aimed to answer several questions; one of them was about the best organisation that led to faster menu search. Therefore, three menu organisations were tested: alphabetic, functional, and random. The menu used in this study consisted of 18 items that were arranged vertically; there were horizontal lines dividing items into groups with no more than four items in each group. In the initial search, users were faster in searching alphabetically organised menu followed by functional and random menu. After obtaining enough practice in using the menu, the difference between all organisations is negligible. And all organisations become the same as a user needs just a single saccade to find an item.

Likewise, Bailly et al. (2014) conducted an experiment to study the effect of several factors, including menu length and menu organisation, on user performance during searching linear menus. Concerning menu organisation, three organisations were tested (alphabetical, semantic and unordered). In the semantically organised menu, the items were distributed in semantic groups. Each semantic group had four items. The validity of the semantic groupings was verified by conducting a semantic relevance study in which participants were asked to rate the semantic relatedness of 120 pairs of items. There was a horizontal separator line that separated each semantic group from others. It was found that participants searched the menu and selected the targets faster in the alphabetical and semantic organisation than in the unordered organisation.

Also, some studies investigated the interaction between the semantic organisation of menu items and other design factors. For example, Halverson and Hornof (2008) studied the effects of semantic grouping and visual cues of semantic relation on visual search. They conducted an experiment. Participants were presented with six structured layout schemes; three variables were controlled in the layouts: the semantic organisation of items in the groups, the existence of group labels and the colouring of the background. Their results indicated that people take advantage of the semantic contents of the words in the interface to help them to find the target during their visual search. Therefore, people search the semantically cohesive groups faster as they can assess the semantic relevance of all words in the group by one fixation, while there is a need to fixate more in a non-semantic group. Regarding the group labels, it was found that there was no difference in selection time between the two conditions of the presence of group labels or not.

Additionally, Brumby and Zhaung (2015) investigated whether the visual grouping cues play a role in speeding up finding items in menus and whether these important grouping cues can delay finding items in menus when applied inappropriately. They conducted an experiment that involved searching for a known item in menus. These menus were different in three things; the use of visual grouping cues (line boxes to group items), the semantic organisation of items in the menu and the size of the semantic group (number of items in each group). Their results indicate that the value of the visual grouping cues depends completely on whether the groups are semantically organised or not. Using visual grouping cues when the groups were semantically organised differentiated these groups and led to finding items faster in the menu. In contrast, using the visual grouping cues when the groups were not semantically organised resulted in slowing down searching times. They have concluded that menu items should be visually organised in semantic groups to assist users in finding items in the menu quicker, and conversely, visually grouping non-semantically related items can hinder menu search.

The effect of the semantic relevance of menu items to the search goal was investigated by some studies. For example, Brumby and Howes (2003) conducted an empirical investigation to show that there is interdependency in menu item assessment as proposed by Young (1998). This means the item selection decision is not just dependent on the relevance of this item to the search goal. However, the relevance of other distractor items to the search goal has a strong influence. In their experiment, first, the quality of the items was rated to assess their semantic relevance to the search goal. Then, these ratings were used to manipulate the quality of the distractor items in menus that were presented to the participants. In each menu, there was one good quality goal item. The menus were different in the quality of the distractor items. There were two conditions: low quality or mediocre quality. In the low-quality menu, all the distractor items were rated as bad distractors (very low in semantic relevance assessment), while in the mediocre quality menu, the average rating of distractor items was moderate. It was found that the number of fixated items was more in the mediocre quality menu compared with the low-quality menu. This indicated that during a menu search, the decision to select the target item is not only dependent on the relevancy of this item to the search goal. However, the other assessed distractor items had a strong influence on the decision between continuing the assessment or selecting the item.

Their results empirically advocate the idea of interdependency in menu item assessment.

Definitely, menu semantics is an influential factor in menu search performance. The way in which menu items are grouped and the titles used to name the groups critically influence the menu search performance (Lee & Raymond, 1993). It is commonly assumed that a menu is efficient when its items are organised matching the users' perception of menu semantics (Schmettow & Sommer, 2016). Therefore, menu designers used some methods such as card sorting to elicit users' perception of menu semantics. Users might be different in perceiving the menu semantics and this might justify the outlying performance in a menu search as suggested by Schiller and Cairns (2008).

In the next section, menu search models that were developed to explain and predict the user performance in menu search are reviewed.

2.6 Modelling User Performance in Menu Search

Predictive menu search models are an efficient method to capture scientific knowledge. They synthesise phenomena that are found in different studies. They can inform designers' decisions and help them determine the best design for a menu without the need for extensive user testing (Baily et al., 2014; Baily et al., 2016; Cockburn et al., 2007). The aim of reviewing the existing menu search models is to assess their suitability to predict outlying performance in menu search tasks. Modelling outlying user performance can help in explaining the outlying performance in the menu search and then isolating the outlying performance cases.

The existing models of menu performance can be classified into two categories: mathematical models and cognitive simulation models. Mathematical models are equations that predict the user performance (search time) based on some menu specifications such as menu length, target location and practice. Cognitive simulation models explain user performance by referring to cognitive processes such as attention, perception, and memory (Baily et al., 2014).

Several mathematical models of menu performance were proposed, such as Total Search Time (TST) by Lee and MacGregor (1985) and Search Decision Pointing (SDP) proposed by Cockburn et al. (2007). The mathematical models are less complicated and easier to apply, but they showed fewer details regarding the menu search process (Baily et al., 2014). The mathematical models are primarily concerned with modelling broad classes of users and not modelling individuals or outliers who need to consider variations due to individual differences.

Several cognitive simulation models were proposed to explain the user performance in menu search tasks by referring to the cognitive process of visual search such as EPIC models by Halverson (2008) and Hornof and Kieras (1997; 1999), ACT-R/PM by Byrne (2001) and a rational computational model of menu search by Cox and Young (2004). ACT-R and EPIC models need to hand-code the production rules that control behaviour. A production is *"an if-statement. It describes an action that takes place when the if 'part' (the antecedent clause) is satisfied."* (Brasoveanu & Dotlačil, 2020). Therefore, the possible behaviours are arbitrarily restricted (Chen, 2015).

A more recent menu search model was proposed by Chen et al. (2015) and was based on Machine Learning (ML). This model differs from the previous ACT-R and EPIC menu search models because it does not assume anything about the possible strategies that can be adopted by the users. Therefore, there is no need to hand-code the production rules that predict behaviour. Instead of hand-coding the production rules, their model uses reinforcement learning (RL) to solve the menu search problem. The prediction of this model was tested against previous empirical results that studied the effect of menu organisation and menu length. The model was able to predict the effect of these factors on menu search performance.

The model by Chen et al. sounds promising to help in understanding the role of menu semantics in outlying menu search performance as it can predict the effect of menu semantics on menu search performance. Therefore, to fully understand this model. The underlying theories of this model and how it works are covered in the next section.

2.6.1 Chen et al. Menu Search Model

This menu search model has adopted the adaptive interaction framework by assuming that menu search is logically adapted to the environment of the interaction, the cognitive and perceptual capabilities of the user and the goal of the user to balance between speed and accuracy (Chen et al., 2015).

Adaptive interaction is a theoretical framework that was defined and promoted by Payne and Howes (2013). This framework was designed to explain how human behaviour is naturally adapted. Therefore, by using this framework, researchers and practitioners are now able to justify many interactive behaviours between people and computers (Payne & Howes, 2013).

According to this framework, human interaction behaviour can be explained via four important components: Environment, Utility, Mechanism, and Strategy (see Figure 2.3). Environment means the ecological structure that is experienced by the user. Utility means the goals of the user (what a user wants to achieve). Mechanism means the capabilities of the user, such as the capacity of information processing. Strategy means a sequence of activities that lead to useful behaviour. The strategy component is determined by the environment, utility and mechanism components. Therefore, this framework considers that strategies are developed as a result of logical adaptation to the environment, utility and mechanism components (Payne & Howes, 2013).

This framework was inspired by previous approaches from different disciplines, including Optimal Foraging Theory, Cognitive Game Theory and Cognitively Bounded Rational Analysis (Payne & Howes, 2013).

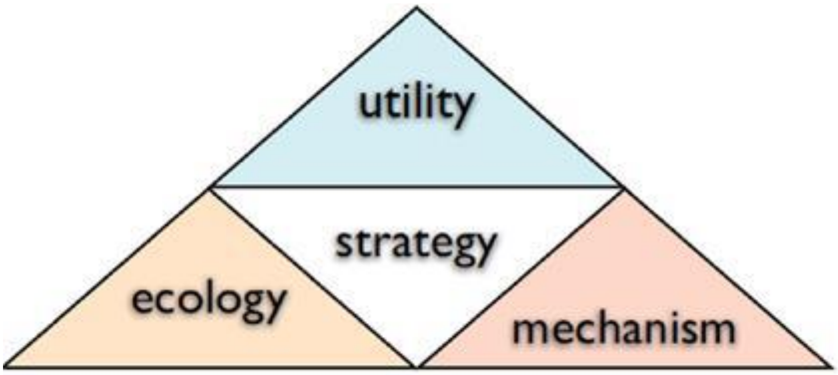


Figure 2.3: The adaptive Interaction Framework (source: Payne and Howes, 2013)

Based on this adaptive interaction framework, Chen et al. (2015) assumed that menu search behaviour emerges as a result of rational adaptation to:

- The environment of the menu interaction, which means in this context menu length (number of items), semantic and shape relevance of menu items.

- The Mechanism, which means in this context the cognitive and perceptual capabilities that allow the user to estimate the semantic and shape relevance of the fixated item and other items.
- The Utility, which means in this context the goal of the user is to maximise balancing between the speed and accuracy during searching a menu.

As this model is based on Machine Learning (ML), Chen et al. (2015) used a reinforcement learning algorithm to implement their model. Reinforcement learning is explained by Sutton and Barto (1998) as "*learning what to do---how to map situations to actions---so as to maximise a numerical reward signal. The learner is not told which actions to take, as in most forms of machine learning, but instead must discover which actions yield the most reward by trying them*". The reinforcement learning algorithm used in this model was a Q-learning standard implementation, see Figure 2.4. The Q-learning was used to solve the menu search problem. Therefore, the menu search behaviour develops by discovering the optimal control policy (control knowledge) that determines when the eye movement and item selection should take place (Chen et al., 2015).

```
        • Initialize $Q(s,a)$ arbitrarily, e.g., set zero for each $(s,a)$ pair.
        • Repeat (for each trial):
              -  Initialise $s_0$, or randomly choose one of the states
              -  Repeat (for each step of the trial):
              -  Choose $a$ from $s$ using policy derived from Q-table (e.g., ɞ-
                 greedy)
              -  Take action a, observe $r$, $s'$
              -  $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \times max_{á} Q(s',a') - Q(s,a)]$
              -  $s \leftarrow s'$
    • until s is a terminal state
```

Figure 2.4: Pseudocode for a Q-learning algorithm (source: Chen, 2015).

To explain their model, they presented an imaginary scenario about a user who has a goal to select the 'Show Next Tab' option that exists in the Safari Windows menu, see

Figure 2.5. First, the user fixates the first top item 'Minimise', then encodes the fixated item, then rejects this item because it is not relevant to the goal, then decides to move their eyes to the next group of items that starts with 'Show Previous Tab', then notices that the currently fixated item 'Show Previous Tab' is similar to the target and notices that the next item in the periphery vision is exactly same the shape and length of the target, then decides to move the eyes to the next item 'Show Next Tab', then makes sure that the fixated item is the target and selects it. As modelling aims to predict this behaviour based on theoretical assumptions, this model does not aim to model how people learn particular menus and the position of particular items. Instead of this, this model aims at modelling menu search tasks generally. Therefore, this model is required to learn from experience the optimal way to search for new target items in new menus that have never been seen before (Chen et al., 2015).

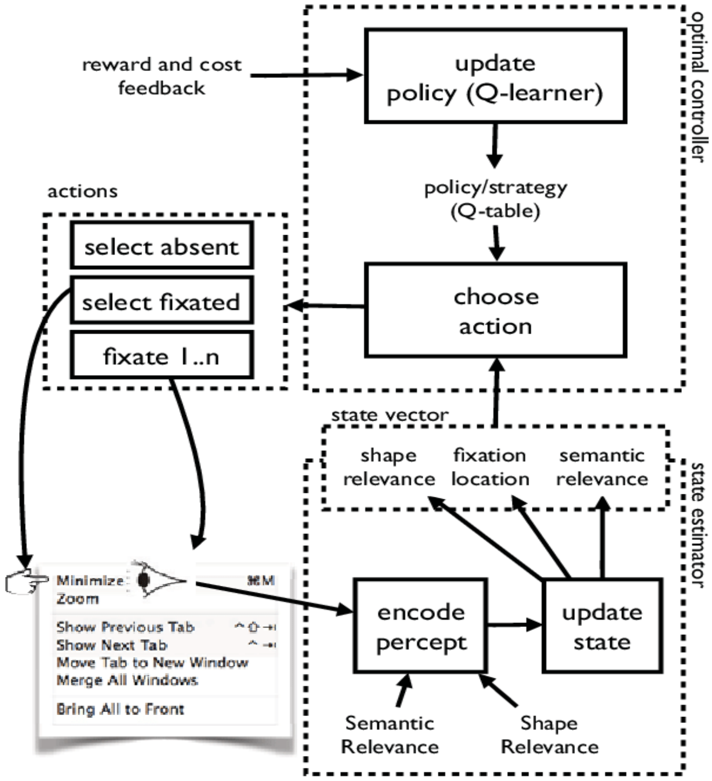


Figure 2.5: An overview of Chen et al. menu search model (source: Chen et al., 2015).

According to Chen et al. (2015), two approaches are used in this model to achieve the aforementioned goal. These approaches are state estimation and optimal control. Therefore, there are two important components in this model; state estimator and optimal controller, see Figure 2.5. The state estimator is responsible for encoding the perception of semantics and shape relevance of the fixated item. This function is constrained by the cognitive and visual capabilities of a human. The optimal controller is responsible for choosing the action (fixate another item, select item, or stop searching (exit menu)).

As illustrated in Figure 2.5, this model works as follows: first, the model fixates the external representation of the presented menu. Then, the state estimator estimates the semantic and shape relevance of the fixated item and encodes this estimation and updates the state vector with this estimation. The state vector has three elements, one for the semantic relevance, one for the shape relevance and one for the location of the current fixation. In the beginning, the state vector items are null, and after every fixation, the state vector is updated by the estimation of the semantic and shape relevance as well as the location of the fixation. After encoding the estimation of the currently fixated item in the state vector, one action (select item, fixate another item or exit) is chosen by the optimal controller based on two things; the state estimation and the Q-table (state-action values). The state-action values (control knowledge) were acquired (learned) by using the reinforcement learning algorithm (Q-learning) (Chen et al., 2015).

The aforementioned description is a brief description of the theory. Many details about how the state estimator and optimal controller work were presented in Chen et al. (2015).

Chen et al. (2015) summarised the assumptions of their model in Table 2.2.

Assumption	Description
Utility	Utility = 10000 * correct -10000 * error - time where correct and error are Boolean variables, and time is in the unit of millisecond
Ecology	Menus have a distribution of length and group size. Menu items have a distribution of semantic relevance and length/shape.
Mechanism	People can estimate the semantic relevance of the foveated menu item. They can also estimate the shape/length of items in the periphery, although acuity decreases with eccentricity.
Strategy	A strategy (or policy) for menu search is optimised to Utility, Ecology and Mechanism assuming a state space that consists of the relevance vectors and the fixation.

Table 2.2: Assumptions summary for the adaptive menu search model (Source: Chen et al., 2015).

Chen et al. (2015) tested the model’s prediction on commonly used application menus (Apple OS). The task was searching for specific target items in a vertically arranged menu. The menu organization was altered. Menu items can be unorganised, alphabetically organised or semantically organised. They determined the ecological distributions of a menu search environment using Apple OS applications menus. They determined the ecological distribution of menu length, item length and semantic group size. Also, they determined the ecological distribution of semantic relevance of menu items. After that, they trained the model until performance plateaued (taking 20 million trials). On every trial, the model was trained on a menu created by sampling randomly from the defined ecological distributions of shape and semantic relevance. The obtained optimal policy through this training was then used to predict the menu search performance. To accomplish this, they tested the model using 10000 newly generated menu samples and recorded their performance. They found that the model was able to predict the effect of menu organization on search duration and gaze distribution.

To sum up, this model seems suitable to study outlying menu search performance as it can predict the effect of menu semantics on menu search performance. Therefore, it might help in understanding the role of the perceived menu semantics in outlying menu search performance.

As this menu search model is based on ML, it requires to be trained in a menu search environment. To prepare a menu search environment, it is required to collect the semantic similarity ratings of menu items. These ratings are used to construct training menu samples that are used to train the model. Therefore, In the next section, the methods used to collect the semantic similarity ratings of menu items are presented.

2.6.2 Collecting Semantic Similarity Ratings of Menu Items

To identify the semantic relationship between menu items, two methods can be used: pairwise similarity ratings and card sorting. These two methods are described and compared in the following sections.

- ***Pairwise Similarity Ratings***

This method is thought to be the gold standard method for generating a similarity dataset (Lantz et al., 2019). It involves asking participants to rate the similarity of each possible two items combination. It was used in many studies that investigated the semantic similarity in psychology and linguistics, such as Miller and Charles (1991), Resnik (1999), and Charles (2000). Also, it was used in some previous studies in HCI to collect the user assessment of the semantic similarity of different menu items. For example, Bailly et al. (2014) assessed the validity of menu semantic groupings by asking 7 participants to rate the semantic relevance (0 to 100) of 120 pairs of menu items. The order of pairs was scrambled, and the participants had no idea whether these pairs came from a semantically organized menu or not. Additionally, Chen et al. (2015) ran a study to collect the semantic relevance ratings. They presented 64 pairs of menu items and asked 31 participants to rate the likeliness that two menu items were found together on a menu.

- ***Card Sorting***

Card sorting was developed by psychologists as a method to understand how individuals organize and categorise their knowledge (Wood & Wood, 2008). Card sorting has been applied in various fields such as designing a user interface, knowledge elicitation, requirement engineering, market research and web design (Schmettow & Sommer, 2016). It is a common method in HCI research and practice. It involves requesting participants to sort a set of items into groups of items that are similar in some way. Card sorting has been used as a technique for organising the information system contents in a way that reflects the user's expectations. It has been used since the early 1980s (Albert & Tullis, 2013). For example, Tullis (1985) used the card sorting technique to organise functions in menus of a mainframe operating system.

In the process of designing menu structures, it is commonly assumed that a menu is efficient when its items are organised matching the user's mental model of this menu’s domain. Therefore, usability designers frequently use card sorting to elicit such mental models (Schmettow & Sommer, 2016). Basically, card sorting assesses the extent of the perceived semantic similarity within a group of items. So, the researchers then group semantically associated items together to construct a structure for these items (Schmettow & Sommer, 2016).

According to Spencer (2009), there are different methods of card sorting; open or closed, team or individual, manual or with software. In closed card sorting, the participants are given predetermined labels and asked to sort the cards according to these labels. In open card sorting, the participants are responsible for sorting the cards into groups and assigning a label for each group. Open card sorting is more common than closed card sorting. This is because open Card sorting yields more information about the groups created by the participants along with the cards under each group, while closed card sorting helps in identifying where the cards would be placed. The participants in card sorting can be participated in a team or individually. In team card sorting, the participants discuss with each other what the groups should be created and where the cards should be placed. In individual card sorting, each participant sorts the cards separately. Individual card sorting is easier to manage and allows for collecting more diverse responses Card sorting can be done manually or with software. In manual card sorting, physical cards are put on a table, and participants are asked to group similar cards next to each other and then gradually group them in piles. In software-based card sorting, the cards are displayed on the screen and participants are asked to drag these cards around into categories. The software-based card sorting is a one-step process as no need to process and analyse the results using another tool. Moreover, using online card sorting allows running the Card sorting remotely.

Many studies in the literature assumed that matching the user's mental model is a requirement for good usability. For example, Faiks and Hyland (2000) used card sorting to organise the content of the online digital library of Cornell University. They found that card sorting was an effective and useful method for gaining user insight into organisational groupings before designing the system. There were also studies that validated card sorting by showing the improvement in usability when the system structures match the card sorting data. For example, Nakhimovsky et al. (2006) reorganised the frequently asked questions list by using expert card sorting. They found that the reorganised list lowered the task completion time by one third, and errors and give-up rates were decreased by half.

- ***Card sorting vs pairwise similarity ratings***

According to Lantz et al. (2019), card sorting is much faster and requires less cognitive effort compared with pairwise similarity ratings. However, card sorting produces binary data. So, important data might be lost, such as the degree of similarity or dissimilarity between stimuli. This is especially true in individual card sorting. However, in aggregate card sorting, many participants complete the card sort and that should produce a dataset of similar quality to that produced by the pairwise similarity ratings method.

On the other hand, the pairwise similarity ratings method is yielding more data than card sorting as it requires making a comparison using a specified scale (e.g., 0 to 10), while card sorting offers binary data. However, the pairwise similarity ratings method is time-consuming and requires a high level of concentration (Lantz et al., 2019).

There have been previous studies in different domains that compared card sorting with pairwise similarity ratings. Van der Kloot and Van Herk (1991) examined card sorting and pairwise rating of personality data and found high correlations between these two methods. Dwyer (2003) assessed the pairwise similarity ratings and card sorting of alcohol expectancy and found a moderate correlation between these two methods. Lantz et al. (2019) found a moderate correlation between the pairwise similarity ratings and card sorting of mental disorders. However, all these studies were in a very specific area and may not apply to HCI. Therefore, it is interesting to collect the semantic similarity of menu items using both methods and check whether they introduce different features on the data. This is important when populating a model based on this data.

- ***Analysing card sorting and pairwise similarity rating data***

Both card sorting and pairwise similarity rating produce a similarity matrix that represents similarity degrees between any two items. The similarity matrix can be then analysed using clustering techniques such as hierarchical clustering and multidimensional scaling (MDS). Hierarchical clustering and multidimensional scaling are statistical techniques that are used to simplify the complex data and organise it in a visual representation to help in the investigation of *"the underlying relational structures"* of how the stimuli were perceived by participants (Lantz et al., 2019).

According to Albert and Tullis (2013), hierarchical clustering analysis and multidimensional scaling (MDS) are two useful statistical methods to analyse the similarity matrix generated by card sorting. The hierarchical cluster analysis builds a tree diagram that shows the cards grouped by most participants in the study in close branches. The multidimensional scaling (MDS) creates a map that shows the distance between all pairs of cards.

2.7 User Performance and Individual Differences in HCI

Individual differences might play a role in outlying user performance in usability testing as it was found by early work on individual differences in HCI that user performance can be varied significantly due to individual differences. For example, Egan (1988) reviewed several studies that collect the task completion time metric on different tasks such as text editing, information search and programming. He found that for a small sample size (10 to 30 participants) in the same text editing task, the difference between the highest completion time and the lowest completion time was 5:1. This variance in performance is not because of differences in the given task but because of differences between individuals. He presented five categories of individual differences in HCI, including technical abilities, age, experience, domain-specific knowledge and personality.

Dillon and Watson (1996) stated that differential psychology could be related to some HCI problems. Also, they argued that the user analysis in HCI could use differential psychology to understand user interaction behaviour. In their review, they discussed the main types of differential psychology along with the previous experimental efforts. The discussed types of differential psychology were cognitive psychology, personality and cognitive style, psychomotor differences and skills acquisition.

In the following sections, some previous works that investigated the effect of individual differences in user performance in usability tests are presented based on the factor studied.

- **Age**

Sonderegger et al. (2016) stated that when it comes to age-related issues, a variety of functional impairments have been cited, including bad eyesight, loss of hearing, a decline in manual skills and deteriorating in memory performance (e.g., Kroemer, Kroemer & Kroemer-Elbert, 2001; Matthews et al., 2000). So, when interactive technology is used, these multiple age-related changes at the cognitive, perceptual, and motor levels are very relevant. In addition to different declines in functional abilities, users' attitudes about technology are influenced by their age. When compared to younger users, older adults exhibit fewer positive attitudes toward new technologies (Chua et al., 1999; Dyck & Smither, 1994). Thus, Sonderegger et al. (2016) studied the effects of age in usability testing. They found that younger participants obtained better performance scores in task completion time compared with older participants. Therefore, they concluded that it is important to consider this factor in usability research and practice as differences in age affect speed and accuracy. This might suggest that age is a very likely factor in outlying performance in usability testing. However, controlling for age did not remove outlying user performance, as found by Auskerin (2012).

- **Personality**

According to Burnett and Ditsikas (2006), a usability testing session is fundamentally a social interaction between the participant and the moderator(s), as well as, to some extent, the participant, and the computer system. As a result, it is quite likely that the personality of the participant will have a significant impact on the whole experience**.**

Therefore, they investigated the role of the participant's personality in usability testing. They conducted a usability testing study and recruited 10 participants, five with high extraversion and five with high introversion, based on a Myers-Briggs test. Myers-Briggs test is a common and widely used test that measures introversion and extraversion, but it has little academic credibility (Stein & Swan, 2019). The participants carried out several tasks on a commercial website and were asked to think aloud while doing the tasks. It was found that extraverts found 40% more usability problems compared with introverts. Also, they took a long time to complete the session.

Similarly, Alnashri et al. (2016) found that test participants' personality dimensions are influential factors in usability testing results. They found that extroverts took more time in doing the task and made more mistakes compared with introverts.

Recently, Schmidt et al. (2019) investigated the impact of the participant's personality on quantitative and qualitative metrics in usability testing. They conducted a usability study with a website, including several tasks. They collected different quantitative and qualitative data. They measured the participant's personality using the big five model, also called the OCEAN model (openness, conscientiousness, extraversion, agreeableness, neuroticism). They found that the personality correlated with some of the inspected usability metrics. There was a significant, moderate and positive correlation between time and extraversion.

All the aforementioned studies suggested that the participant's personality influences the participant's performance in usability testing. More specifically, the extraversion dimension was found an influential factor in the time-on-task metric. Extroverts are slower in completing a task than introverts. This might be because extroverts tend to be optimistic, motivated, and less anxious about problems when doing tasks, while introverts tend to be more anxious if they face a problem and therefore, they move on to the next task (Alnashri et al., 2016).

Based on the abovementioned studies, it might be that participants' personality dimensions play a role in outlying user performance in usability testing. For example, it might be that the conscientiousness dimension has a role in outlying user performance as conscientious persons are more dutiful and accurate and that might affect their time-on-task.

2.8 Conclusion

Outlying performance is potentially a feature in usability testing and not just a statistical nuisance. There is a lack of studies that investigated outliers in usability tests. Therefore, there is a need to investigate this problem and find out the reasons behind it.

This thesis is about outlying performance in menu search. Menu semantics is a key factor in menu search performance. Previous work suggested that the perceived menu semantics plays a role in outlying menu search performance. The menu search models can help in understanding the role of the perceived menu semantics in outlying menu search performance. Using such models needs collecting semantic similarity ratings of menu items from human participants. There are ways to collect the semantic similarity ratings, such as pairwise similarity ratings and card sorting.

Additionally, outlying performance could be caused by specific individuals. Therefore, there is a need to consider individual differences that might be relevant to outlying performance in HCI.

Chapter 3

Study 1: Investigating Outlying Performance in Usability Testing Practice

3.1 Introduction

To motivate investigating outlying performance in general, it is important to check whether usability practitioners take outliers seriously and whether they have ways of dealing with them. Moreover, it is important to know how practitioners interpret outliers in usability tests as this might help in framing the findings of the subsequent studies in this research.

In fact, no previous works investigated how usability practitioners interpret and treat outliers in usability tests. Folstad et al. (2012) investigated how practitioners treated single-user problems in usability tests. Outliers can be considered a special case of single-user problems. However, in their work, they just investigated treating these single-user problems, but they did not investigate the possible causes of these single-user problems.

Therefore, this exploratory study aimed to investigate how outliers are interpreted and treated in usability testing practice.

3.2 Method

3.2.1 Design

This study aimed to gain knowledge on how practitioners interpret and treat the outlying performance cases in usability tests. Interviewing was considered the most appropriate method for determining how practitioners interpret and treat outliers when they occur. This was because there is no existing account of how outliers are dealt with though it is likely that they nonetheless have to be addressed. Interviews allow for probing this specific issue but also following up on particular topics depending on what the interviewees said.

Interviews, of course, rely on self-report and may therefore not capture the actual behaviours and attitudes of practitioners. Observations would overcome this by allowing the researcher to see what practitioners actually do, but without being able to guarantee when outliers might be observed, this was considered to require an impractical intrusion into the participants' work.

Interviews are defined as a "conversation with a purpose" (Kahn & Cannel, 1975, as cited in Lazar et al., 2017). There are four interview types: structured, unstructured, semi-structured and focus group interviews (Fontana & Frey, 2005). The semi-structured interview allows more flexibility than a structured interview. This is because the structured interview follows rigid scripts while asking questions, and no ways to add more questions that did not already exist in the specified interview script. While in the semi-structured interview, there is flexibility in adding more questions, asking for clarifications, and discussing any issues that concern the interviewee. Therefore, there is a great chance to find interesting information that might lead to more focus and deep understanding. However, more skills are required to control the semi-structured interview. Moreover, analysing the collected data can be more challenging due to the less structure (Lazar et al., 2017). The unstructured interview is similar to a conversation about a specific topic. It allows going into more depth. Questions asked in this type of interview are open, meaning no prior expectations about the answers' format and content. Although unstructured interviews produce rich data that gives a deep understanding of the topic, they are not consistent among participants as each interview has its own format (Preece et al., 2015).

In this study, a semi-structured interview was the most appropriate type of interview to collect the data from the participants as it allows digging deeper when there is a possibility to gain more understanding. Moreover, it enables objective comparison of participants' answers.

After collecting the data from the participants via the interviews, thematic analysis was used to analyse the collected data. Thematic analysis is a widely adopted qualitative analysis method. It provides an accessible and theoretically flexible method to analyse qualitative data. It is used to identify, analyse and present themes that reside within the collected data (Braun & Clarke, 2006). These identified themes are important and interesting to address the research problem (Maguire & Delahunt, 2017). The thematic analysis was the most appropriate method in this study that aimed to investigate how practitioners interpret and treat outliers in usability tests. This was because this method tries to find patterns of experiences of practices and attitudes to those practices, as stated by Cairns and Cox (2008). The thematic analysis was also the most appropriate method compared with other qualitative analysis methods such as grounded theory and content analysis. This was because this study had a specific focus, namely the interpretation and treatment of outliers in usability tests. The grounded theory allows for the emergence of the underlying theory from the collected data (Lazer et al., 2017), it would allow for the emergence of causal accounts around a phenomenon of interest, but it was felt that the focus was sufficiently well defined and therefore, did not require such an open approach to analysing the data. The content analysis involves examining the interview data for usage patterns, and this includes the frequency of specific terms, cooccurrences and other indicators of the significance of different concepts and the association between them (Lazer et al., 2017), but in this study, there was no need to count specific terms and find the relationship between them.

The thematic analysis consists of six phases as developed by Braun and Clarke (2006). These phases are as follows:

1- Getting familiar with the data.

2- Generating initial codes.

3- Searching for themes.

4- Reviewing themes.

5- Defining and naming themes.

6- Producing the report.

The aforementioned six phases are not linear, and it is possible to move back and forth when there is a need. So, thematic analysis is an iterative process that needs to iterate generating the codes and themes several times until it is ensured that all the identified themes are covered all relevant data (Braun & Clarke, 2006).

3.2.2 Participants

To recruit participants for this study, I searched for organisations in Saudi Arabia that practice usability tests, such as software companies and innovation and design consultancy companies. I conducted my searches via Google, Twitter and LinkedIn. Several organisations were found. After contacting these organisations and asking them to provide me with the contacts of their practitioners, a list of candidate participants was created. The candidate participants' qualifications, certificates and experiences were checked by visiting their LinkedIn pages. This was to ensure recruiting participants who have strong profiles, which were indicated by their experiences, qualifications, or certificates. Thirteen participants were recruited, seven males and six females, their ages ranged from 24 to 40 with a mean age of 28.3 years (SD = 4.5), and their working experience as UX practitioners ranged from 1 year to 14 years with an average working experience of 4.3 years (SD = 3.4), see Appendix A.2. To vary the collected data, a maximum of two participants were interviewed from the same organisation. The corporate UX maturity of the participants' organisations ranged from stage 3 to stage 7 based on Nielsen's usability maturity model, see Appendix A.2

In Nielsen's usability maturity model, the development of the UX processes in the organisations usually follows the same phases starting from hostility toward usability until extensive focus on user research (Nielsen, 2006), see Appendix A.3. This model was chosen to assess the corporate's maturity because it is lightweight and provides detailed English documentation (Salah et al., 2014). So, the participants were asked to read the documentation of this model and then assess the maturity of their organisation by matching their organisation with one of the maturity stages 1-8.

3.2.3 Materials

The interview questions followed the commonly used sequence mentioned by McCartan and Robson (2016); introduction, warm-up, main body of interview, cool-off and closure, see Table 3.1.

Interview Questions
Introduction • Introduce myself. • Explain why I am doing the interview. • Reassure the interviewee regarding the ethical issues. • Ask for permission to record the interview.
A warm-up • Would you like to tell me a little about yourself? • How many years have you been working in this field of "usability testing"?
The main body of the interview • Do you carry out the usability testing? • Could you tell me how you do the usability testing? • Do you use measurement during usability testing? • What do you measure? • After collecting the data, what do you look for in the collected data? • If you find data that is different substantially compared with the rest of the data "outlier", what are you going to do with this data? • What are you going to do about this user? • What do you think is the reason behind the occurrence of outliers in usability testing?
A cool-off • Do you enjoy working in this field? • How do you develop your skills in this field?
A closure • Thank the interviewee for his/her participation. • Switch off the recorder. • Indicate that the interview has ended.

Table 3.1: The interview Guide.

As the questions in the main body of the interview are considered the essence of this study, they were developed carefully and ordered logically to get a better insight. They were simple and unbiased. The questions did not lead participants as they did not indicate that outliers frequently occurred in usability testing. Also, they did not imply that having outliers in usability testing is problematic. They just focused on asking about different data and unusual behaviour and how to deal with outliers. The "outlier" term was explained as a test participant who took a long time to complete the task compared with other test participants.

I started with broader questions, and then I moved to more specific questions that focused on the goals of this study. The broader questions helped to understand the business process of the participants' companies and how they conduct the usability testing in their companies, and what sort of measurements they use. This helped in understanding and analysing the answers to the later specific questions that focused on the goals of this study.

3.2.4 Procedure

Some interviews were face-to-face, and some were via telephone. Telephone interviews were conducted only when a face-to-face interview was difficult to arrange because participants were geographically distant.

At the beginning of each interview, participants were welcomed and made to feel at ease as it is important to establish a good rapport with participants in the first few minutes of an interview (Kvale, 2007). Then, they were introduced to this study by briefly explaining the aim of the study. Next, the information sheet and consent form were given to the participants to read and sign, see Appendix A.1. This was to make sure that the participants understood the purpose of this study and to reassure the participants regarding ethical issues such as anonymity. After reading the information sheet and signing the consent form, the participants were asked for permission to record the interview. Then, the audio recorder was switched on, and the interview started.

Some *ad-hoc* questions were asked based on the previous answers. After asking all questions and making sure that all aspects had been covered, the participants were thanked for their participation in this study, and then, the recorder was switched off.

After finishing each interview, I tried to take a few minutes to write notes related to the session as suggested by Kvale (2007). I transcribed the interview by writing the questions and the full answers. I tried to transcribe each interview within 24 hours. This allowed me to write my comments that might need to follow up with the participant to ask more questions. Some interviews were in the Arabic language. Therefore, the answers were translated into the English language.

3.2.5 Data Analysis

As mentioned before, the thematic analysis was adopted in this study to analyse the collected data. To identify the themes in this study, three approaches were used. First, the inductive ‘bottom-up’ way was used to find the themes within the data. In this way, the identified themes are very related to the data themselves (Braun & Clarke, 2006). The second approach was semantic coding. The semantic approach focuses on the semantic or explicit meaning of the data to generate the codes that formed the themes and does not look for any interpretation beyond the collected data (Braun & Clarke, 2006). The third approach was using the essentialist/realist analysis. The essentialist/realist analysis means focusing on and presenting the participants’ experiences, meanings and reality (Braun & Clarke, 2006).

Braun and Clarke’s six phases guide of the thematic analysis were followed. First, to get familiar with the collected data, the audio recording of each interview was heard two times, one to transcribe the interview and another one to review the transcription. Although the transcription process might be time-consuming, it can be a great way to start familiarisation with data (Riessman, 1993, as cited in Braun & Clarke, 2006). After that, the transcriptions were read and re-read several times to be fully familiar with all issues in the collected data.

After getting familiar with the collected data, the code generation process was started. In this phase, each segment of data relevant to the research question was coded. Open coding was used, which means that codes are developed and changed through the coding process. Each piece of data was given a labelled code that was added to the transcriptions document as a comment beside the highlighted interesting data. Examples of the generated codes were usability testing methods, usability metrics and outliers’ treatment strategies.

Then, all the generated codes were grouped and collated to search for potential themes. The similarities and differences among the generated code were identified. This helps in making sure that the generated codes and their corresponding data extracts could form themes and sub-themes. Four themes were identified. Examples of these themes were the flexibility of usability testing practice and the treatment of outlying performance cases in usability testing.

These four identified themes were reviewed to make sure that for all themes, there were enough data to support them and that there is consistency in the data that belonged to the same theme. Moreover, the identified themes were reviewed to make sure that they reflect all interesting issues in the entire collected data.

Once satisfied with the identified themes, each theme was defined and named for the analysis. Producing the report phase was done concurrently with the defining and naming themes phase because defining and naming themes phase necessitates writing about each theme along with the corresponding data extracts (see section 3.4). The four themes were reported in a narrative way that ultimately presented the findings that achieved the aim of this study. In the next section, these identified themes, along with the corresponding data extracts, are presented in detail.

3.3 Results and Discussion

Four main themes were identified during the data analysis. These themes were named as follows:

1- The flexibility of the usability testing practices.

2- The awareness of outlying user performance in usability tests.

3- The qualitative interpretation of the outliers in usability tests.

4- The treatment of the outlying user performance cases in usability tests.

These four themes are described in detail in the following sections.

3.3.1 The flexibility of the usability testing practices

It was found that different practitioners adopted different methods of usability testing and used different usability metrics. Moreover, the same practitioner did not necessarily just use one method and one metric but would vary them according to the context of the project.

Some practitioners adopted moderated usability testing and described how it is carried out. In the moderated usability testing, the moderators are present with the test participants, introducing the session, explaining the testing scenario, answering their questions, listening to their feedback, and taking notes. As a practitioner said:

> *If I am talking as a moderator, then I start by introducing the tasks, what I would like my users to do. Usually, I use the thinking aloud technique. So, I ask the user to illustrate anything that he is trying to do. (..) If the users are stuck on something, I will give them some time. I observed their facial expressions and the way they deal with the system, and I gave them some time. Only once I see them that they are hopeless, they are just stuck there. I tell them, okay, a hint would be here. (P1).*

The moderated usability testing can be conducted outside the organisation:

> *"There are many ways to do the usability testing, but our approach in usability testing is moderated and done outside the company" (P10).*

Some practitioners said that they adopted unmoderated remote usability testing, which does not need a moderator because it is done remotely using some sort of online usability testing tools:

> *"We do not have a usability lab, but I use online tools such as Crazy Egg and Usabilla." (P4).*

The practitioners pointed out that usability testing can vary in complexity, and therefore, the lab needed to conduct it:

> *"The testing can be simple (no need for UX lab) or can be complex (needs special software and hardware such as eye-tracking and research systems)" (P6).*

The practitioners indicated that usability measurements could be driven by constraints coming from very different sources:

> *"The used measurements depend on our goals of the usability testing." (P4);*

> *"The usability measurements are based on the goals of the project which depend on the objectives of the business and the hopes of the user." (P7);*

> *"We focus on the concern of the business owners, for example, if they want their product to be fast, easy to use." (P8).*

Additionally, they mentioned that the adopted usability measurements are affected by the organisation's resources:

> *"We use quantitative and qualitative measurements, but we heavily depend on the qualitative measurements because the quantitative measurements need a large number of the participants, which is sometimes hard to achieve." P(12).*

The practitioners mentioned various usability measurements. They measure task time, task success and satisfaction. They record the screen, mouse click, and keystroke of the keyboard. As stated in the following data extracts:

> *"We do use the measurement during the usability testing. We have to test the satisfaction rate. We use the time and the success rate of the task. Also, we measure the mouse click, screen recording, keystroke of the keyboard, the mouse travel, the scrolling time, or the number of scrolls. All of these taken into consideration" P(2);*

> *"We use quantitative measurements such as success criteria and time on task." P(5);*

> *"Recording the screen, mouse clicks, taking notes about the user interaction. Eye-tracking is not always accurate. The best thing is to see the user and let them talk during the usability testing" P(7);*

> *"We measure the time spent to achieve the goal and if he did the task successfully. Also, it is necessary to measure the satisfaction level by using a survey." P(10).*

This variety in usability testing methods and measurements indicates that usability testing practice is flexible. This flexibility in usability testing practice is considered one of the great features of usability evaluation as no compulsion on a specific type of evaluation method (Albert & Tullis, 2013). This flexibility is needed to fulfil the requirements of the business sponsor (who funds the usability study) and to reflect the resources of the organisation (who conduct the usability study).

3.3.2 The awareness of outlying performance in usability tests

It was found that some practitioners explicitly mentioned that they always have outliers in usability testing when they were asked in the interview about dealing with data that are different substantially compared with the rest of the data (outliers). They said:

> *"I carried out the usability testing nine times, and every time I found outliers, especially in time on task" P(2);*
>
> *"We faced outlying performance many times" P(7);*
>
> *"Yes, this always happens" P(10);*
>
> *"Definitely, we always have outliers" P(12).*

The awareness of the negative impact of an outlier on the whole data was not mentioned by most of the practitioners. This might be because I did not ask about this issue directly. However, just one practitioner pointed out this issue and said:

> *"We will do a geometric mean which is a calculation that reduces the effect of the outliers because one outlier might ruin the whole data" P(5).*

The awareness of the practitioners regarding the regular occurrence of outliers in usability testing confirms the observation made by Schiller and Cairns (2008) when they stated that there is always one user in usability tests who is substantially slower than others.

3.3.3 The qualitative interpretation of the outliers in usability tests

The practitioners provided several interpretations of outliers in usability testing. The majority of the practitioners were strongly convinced that people are different in everything. Therefore, they interpreted the outlying performance cases as a result of individual differences in age, personality, experience, technical background, etc. As the practitioners interpreted:

> *"I believe the participant's experience is one of the reasons behind the occurrence of the outliers." P(2);*
>
> *"People think and behave in different ways." P(3);*
>
> *"Some people get nervous when you ask them to do the task, and that affects their performance. Also, their technical backgrounds, ages and personal circumstances can affect their performance" P(4);*
>
> *"People are different in their mental models, ages, technical levels, experiences with this kind of application and personality (open person or not) will affect." P(5);*
>
> *"People are different in everything (culture, education and experience)" P(7);*
>
> *"Maybe the willingness of the user to participate in the usability testing, maybe the user has personal circumstances" P(8);*
>
> *"Maybe the personality of the user and maybe the user has personal circumstances such as illness" P(10);*
>
> *"Maybe the mental model, the previous experience and the educational level." P(11).*

This interpretation of outlying performance is plausible. It is in line with what was stated by Egan (1988) that individual differences lead to considerable variance in performance among participants in HCI studies. And he concluded that the reasons for such variations seem to be related to variables including technical abilities, age, experience, domain-specific knowledge and personality. Age was found as an influential factor in useability research as it affects the speed and accuracy of the participants (Sonderegger et al., 2016). Also, test participants' personality dimensions are influential factors in usability testing results (Alnashri et al., 2016; Burnett & Ditsikas, 2006; Schmidt et al., 2019).

Another interpretation of outlying user performance in usability tests was an artefact of the testing situation. The practitioners pointed out that moderators and their instructions during the usability testing session might affect the participants' performance as the practitioners stated:

> *"Sometimes even if we tried to keep all the factors the same across the session it's not always possible, sometimes you said a word and this word confused everything in the session" P(1);*
>
> *"Maybe the moderator did not do well during the session" P(10).*

When usability tests were conducted online, outlying performance may happen because participants leave their laptops and take a break as a practitioner said:

> *"Maybe the user leaves his laptop and does something else" P(4).*

This interpretation is in line with an interpretation mentioned by Albert et al. (2010) that extremely long task times in online usability tests usually indicate that participants have taken a coffee break or even have gone home for the evening.

The practitioners also attributed outlying user performance to poor understanding of the instructions or not following the instructions given in the usability testing session:

> *"It might be that the participant did not follow the testing instruction and tried to discuss the problems during the session." P(9);*
>
> *"Maybe he has a poor understanding of the instructions." P(10).*

Recruiting nonrepresentative test participants was one of the provided interpretations of outlying user performance in usability tests. Some practitioners stated that there is a possibility that the test participant was mistakenly recruited. Therefore, the outlying participant was not a representative of the target audience and most probably that his/her interaction with the system was different. As stated by the practitioners in the following data extracts:

> *"Maybe the user itself, when I recruited this user was not from the target of the usability testing" P(1);*
>
> *"The user is not from the target" P(5).*

These interpretations of outlying performance cases as an artefact of the testing situation are possible. However, the outlying performance cases that are interpreted as a result of these aspects should not be considered as stated by Nielsen (2000), *"there is always a risk of being misled by the spurious behaviour of a single person who may perform certain actions by accident or in an unrepresentative manner".*

A few practitioners indicated that outlying performance is not always caused by the user itself. It might be caused by usability problems in the tested system. As stated in the following data extracts:

> *"Some time the interface has a problem, and it does not provide informative feedback" P(5);*
>
> *"Not always the problems come from the user, maybe something is missing in the prototype" P(10).*

To sum up, the above-mentioned interpretations indicate that the interviewed practitioners in this study tended to link the poor performance to the test participants instead of linking that to problems in the interaction with the tested system.

3.3.4 The treatment of the outlying performance cases in usability testing

It was found that the majority of the practitioners claimed that they consider the outlying performance cases, and they do not exclude the outliers unless there are reasonable justifications. They investigate the outlying performance cases, and based on their investigation, they decide on accepting or removing the outliers. As outlined by the practitioners in the following data extracts:

> *"I will dig deeper and investigate the session (...) it is our responsibility to accommodate all users." P(1);*
>
> *"We will not ignore the data, and we need to look at his profile (old age, background, experience). Most of the time, I have answers to why this user struggles. Every data is important (...)*

> *our exclusion has to have a strong reason behind it to validate it" P(5);*
>
> *"We do not eliminate the outliers at all. We have to take them into consideration." P(8);*
>
> *"Exclude the extreme cases if there is a strong justification" P(10);*
>
> *"We do not ignore the outliers. We have to look for why this happens" P(12).*

Although they claimed that they always consider outliers and investigate them carefully, it might be that practitioners claimed that because they want to show that they are good practitioners. This can be considered a social desirability bias that occurs when participants respond inaccurately only to be better accepted by others.

Some practitioners think that outlying performance should be considered if it happens in important tasks:

> *"If this outlying performance occurred in an important task, we should consider it." P(3);*
>
> *"I have to check if outlying performance happened frequently and whether it was in critical tasks and then decide to consider this case or ignore it" P(4).*

The practitioners suggested some *post-hoc* strategies to treat outlying performance cases. These strategies are specified after the outlying user performance cases were found.

The practitioners frequently suggested reviewing the notes and the session recording:

> *"I go back first to my notes, then to the recording of the session itself, then to the user itself" P(1);*
>
> *"I have to check my notes, his talk. (..). If this happens again with another user, that means there is a problem in the interface that needs to be fixed" P(7);*

> *"We have to look for why this happens. Usually, we have a video recording for the whole session, so it is clear if the user had a problem during the interaction. Also, we back to eye-tracking data so you notice what they were saying, what they were doing and where they were looking. These three things can indicate what the problem is." P(12).*

This highlights the importance of collecting different types of data during the usability testing session. These different types of data can supplement each other and help in understanding why outlying user performance happened.

Some practitioners suggested discussing with the outliers to understand the problem. As they said:

> *"I must ask him about his interaction" P(7);*
>
> *"We will try to find why this happens by asking the user immediately (after the session by the moderator)" P(3).*

This strategy is in line with a recommendation to use a debriefing session to get more information about what was observed during the testing session (Rubin & Chisnell, 2008).

Some practitioners suggested discussing the other team members to reach a judgment on this problem. As suggested by a practitioner in the following data extracts:

> *"We have to discuss the case as a team" P(3);*
>
> *"I have to consult the team members whether to remove this case or accept this case based on the experience of these members" P(11).*

If the reason behind the outlying performance was not identified, some practitioners suggested doing the test again with different test participants as they said:

> *"Doing the session again with other participants" P(10);*
>
> *"I will try to bring other users for the same test" P(11).*

Some of the above-mentioned strategies, such as conducting a new test, discussing with team members and discussing with test participants, are mentioned by the usability practitioners in Folstad et al. (2012) study to handle single-user problems in usability tests.

These diverse *post-hoc* strategies to deal with outlying user performance indicate that no systematic approach was followed by practitioners to handle outliers in usability tests. This confirms the findings of Folstad et al. (2012), who concluded that no established procedures to handle single-user problems in usability tests.

3.4.5 Other Themes

Other themes were identified during the analysis of the interview transcriptions, such as the challenges that face the practitioners in Saudi Arabia and the Middle East. These challenges were as follows; the difficulty in recruiting the test participants because there are no recruiting agencies like in the US and UK. Also, the businesses are not aware of the need for adequate time to carry out the usability studies. Therefore, they always rush and want to see the results quickly. Another identified theme was the intersection of UX and business, which is explained by the importance of the UX for the business and how businesses will benefit from the UX in increasing their profits. These themes were interesting, but they were not presented in more detail because they were not relevant to the aim of this study.

Overall, when asking usability practitioners about how they conduct usability tests, I, of course, find things that are well-known in the field. The most noticeable theme that emerged was the flexibility that practitioners have to conduct usability tests both to meet the needs of the project and the constraints of their organisations. This flexibility of approach is both a strength of the approach and a rational business decision: an inflexible approach to a single style of usability test might miss business opportunities while also being “overkill” for the project’s goals.

However, this flexibility is perhaps the source of some of the specific challenges of dealing with outliers. While many practitioners were aware of outliers, if not always explicitly, there was considerable freedom in how outlying data was handled. This could be a problem for the outcomes of usability tests. Only because one person took a long time to carry out a task does not automatically mean they are not suitable to be analysed or mean they are not from the target audience. In both situations, it could be considered viable to define beforehand what an atypical user might be like and thus remove all users of that type, not just the ones that come to attention because they produce outlying data.

Indeed, several mentioned strategies suggest a more *post-hoc* approach to outliers: they probe for reasons only after they see the outlying performance. This can lead to a potential confirmation bias: once a reasonable reason has been found, that user could be discounted or excluded from further analysis. However, this might not be valid and might lead to removing potentially important insights from the outcomes of the usability test.

According to the information provided, it appears that practitioners are often looking for a way to remove or exclude outliers. They do not like the outliers, even if they know that they get them all the time. Therefore, they attribute the outliers to all sorts of things and then work out why they should be excluded or replaced.

3.4 Conclusions

This study aimed at investigating how outliers are interpreted and treated in usability testing practice. The key findings of this study are as follows; The interviewed practitioners seem aware of the regular occurrence of outliers, they tend to link outlying performance cases to individual differences instead of linking that to usability problems, and there is no systematic approach to addressing them.

The findings of this study should be considered in light of some limitations that might have affected the validity of these findings. The interviewed practitioners were all working in Saudi Arabia. Moreover, their ages range from 24 to 40, with a mean age of 28.3 years (SD = 4.5). So, the sample was narrow geographically and in terms of age. However, the participants are from different nations and training backgrounds. It would be better in future works to make the sample more diverse and interview practitioners from different countries. It would also be better to set a criterion for selecting candidates, such as selecting candidates who have conducted ten usability tests. This will help in generalising the results to the wider population.

The second limitation is the possibility that practitioners believed, at some point in the interview, that their practices were being evaluated. Therefore, they might answer questions in a way that presents them as professional practitioners. However, I tried to mitigate the impact of social desirability bias by phrasing the interview questions neutrally. For example, the questions did not imply that outliers are problematic in usability tests, and they should be treated carefully. Additional limitation is that the qualitative data analysis (thematic analysis) of what practitioners said about how they interpret and treat outliers in usability tests may be susceptible to researcher bias and misinterpretation. However, to mitigate this, another researcher was asked to work individually to code the reports from practitioners and then discuss them with the main researcher together until they agree on reports about how practitioners interpret and treat outliers.

Chapter 4

Study 2: Investigating the Role of Menu Semantics in Outlying Menu Search Performance

4.1 Introduction

This research was motivated by the work of Schiller and Cairns (2008) who identified outliers in menu search and attributed this to the perceived menu semantics based on modelling work. However, modelling alone is not a substitute for gathering and analysing data of real users searching menus, which offer more definite empirical results that support any claim. Therefore, there is a need to address that limitation by conducting empirical studies that check whether the perceived menu semantics plays a role in outlying menu search performance.

Generally, menu semantics is an influential factor in user performance in menu search tasks, as found by several previous experimental studies such as Card (1982), Halverson and Hornof (2008), and Bailly et al. (2014).

Brumby and Zhaung (2015) warned against poorly organized menus that visually group semantically unrelated items as they can hinder menu interactions. Accordingly, it may be that outlying menu search performance is a result of the poor semantic organisation of menu items, at least as perceived by some users. It might be that during searching a poorly organised menu, the user encounters a semantically unrelated item that informs him that the target is more likely not to be located in that part of the menu although it is located there, and then they decide to move to another part of the menu. So, the decision to move to the other part is a cause of his outlying menu search performance.

Therefore, this study aimed to investigate whether the poor semantic organization of menu items plays a role in outlying menu search performance. I hypothesised that more outliers are found in searching a poorly organised menu that visually groups semantically unrelated items.

4.2 Method

4.2.1 Design

This study hypothesized that more outliers are found in searching a poorly organised menu than in searching a semantically organized menu. Therefore, this study adopted a between-group design with one independent variable, which was a menu organisation (how menu items are organised) and one dependent variable, which was a number of outliers (an outlier is a participant who takes a long time to find and select the target item and then appears as an outlier in the boxplot of the collected menu search time data).

Two conditions of menu organisation (semantically organised menu and randomly organised menu) were identified. In the semantically organised menu, the menu items were grouped into cohesive semantic groups, while in the randomly organised menu, the menu items were mixed and grouped in non-semantic groups.

A between-group design was used in this study to test whether the poor semantic organization of menu items plays a role in outlying menu search performance. The between-group design was the most appropriate design approach because of the need to avoid the learning effects that might occur if the within-group design was used. The learning effect was more likely to occur as participants who completed the first menu search task under one condition will know the task, and that might affect their performance during the next task, which is under the other condition.

4.2.2 Participants

Overall, thirty-nine participants took part in this study, 19 participants in the first condition and 20 participants in the second condition. Their ages ranged between 20 and 22, with a mean age of 20.8 years (SD = 0.77). They were all students in the College of Computer and Information Science (CCIS) at King Saud University (KSU). They were all female as KSU adopts a single-gender education. They were confident in their technical skills in using computers and browsing websites. Participants were assigned to conditions randomly. All participants were asked in person if they were willing to participate in the study.

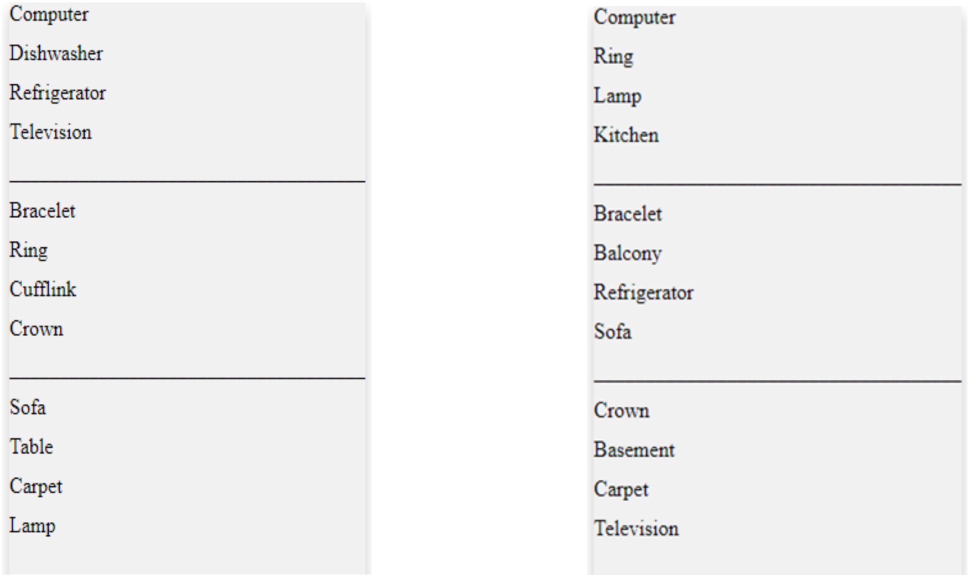
4.2.3 Materials

The menus used in this study consisted of 16 menu items. These 16 menu items are arranged vertically into four groups that are separated by horizontal separator lines similar to traditional linear menus. Each group has four items because four items is the average number of items for each logical group (Bailly et al., 2008).

The 16 menu items were obtained from the database of words created by Yoon et al. (2004). This database was used in previous studies in menu semantics like Bailly et al. (2014), Brumby and Zhaung (2015) and Halverson and Hornof (2008). This database is valuable since it comprises 560 unique words grouped into 105 natural categories.

The 16 menu items belong to four categories: appliance, jewellery, furniture, and room in the house. These categories were chosen because they are similar to some extent to categories found on commercial websites such as Amazon. Four items were selected from each of these categories. Under the appliance category, computer, dishwasher, refrigerator, and television were selected. Under the jewellery category, bracelet, ring, cufflink, and crown were selected. Under the furniture category, sofa, table, carpet, and lamp were selected. Under room in house category, basement, bedroom, balcony, and kitchen were selected.

In the semantically organised menu condition, the 16 menu items were grouped into four cohesive semantic groups, see Figure 4.1 (a), while in the randomly organised menu condition, the 16 menu items were mixed to create four non-semantic groups, see Figure 4.1 (b).



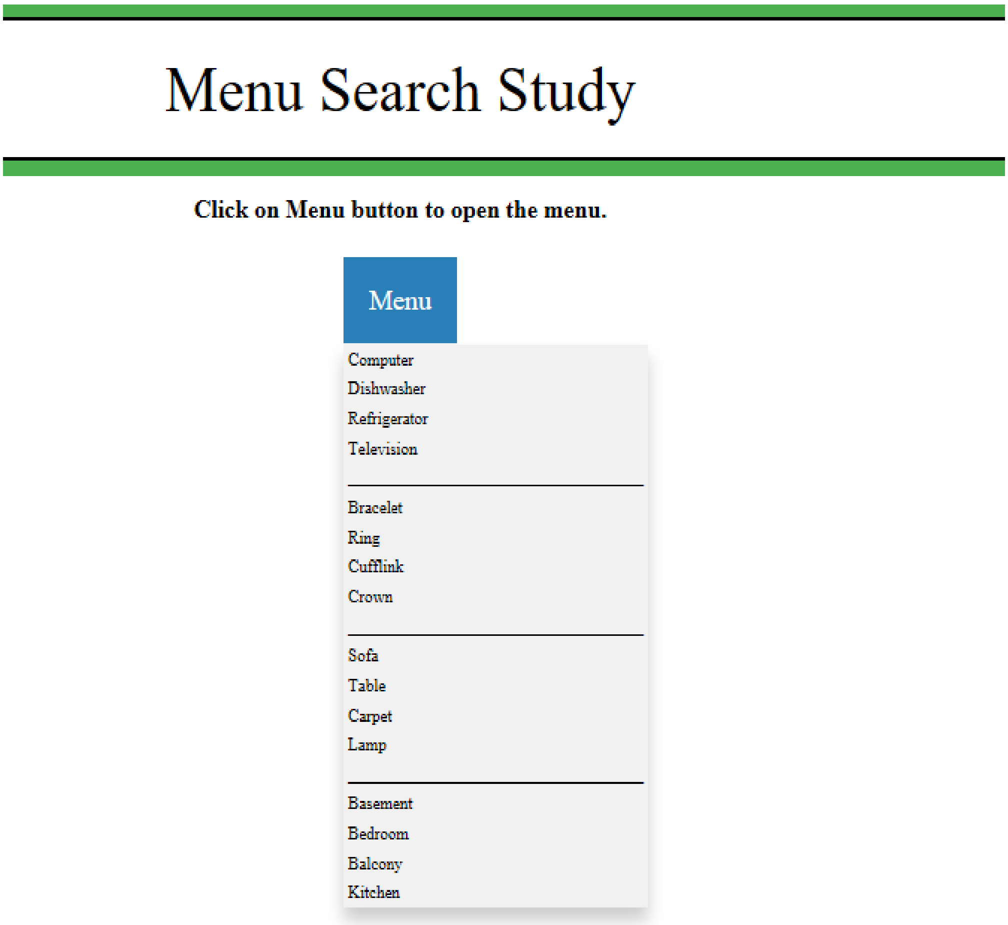
(a) *(b)*

Figure 4.1: (a) A semantically organised menu, and (b) A randomly organised menu.

The menu was displayed on a website developed specifically for this study. Developing the material of this study from scratch allowed customising some design aspects of this study, such as a menu layout. Also, it allowed the recording of participants' performance data in the database of the developed website.

The technologies that were used to develop this website were Hypertext Mark-up Language (HTML), JavaScript, PHP, and MySQL. The HTML was used to create the structure of the website pages. The JavaScript was used program handling the events that happened during interacting with the website. Also, PHP was used to program the interaction with the database of the website. The database of this website was created to store the collected data that was obtained during the experiment, such as participant ID, task start time, task end time, selection time, and selected item.

The developed website consisted of three pages. The first page welcomed the participants and allowed them to enter their ID, and redirected them to the second page, see Appendix B.1. The second page contained a "Menu" button that should be clicked by the participants to display a menu and start the selection task, see Figure 4.2. The third page showed a thank message for participation, see Appendix B.1.



Menu Search Study

Click on Menu button to open the menu.

Figure 4.2: The menu page in the developed website.

Two versions of this website were created to reflect the two menu conditions in this study. These two versions were only different in the organisation of the menu displayed on the second page. In the first version, the menu items were organised in semantic groups. In the second version, the menu items were organised randomly in non-semantic groups.

The two versions of the website were uploaded to a web server to be accessible by the web browser, and that allowed the participants to visit the website and participate in the experiment at the same time.

4.2.4 Experimental Tasks

The main task in this study was to search for a specific item in a menu and select this item. The menu search task starts once the button "Menu" is clicked and the menu appears on the screen. The menu search task finishes when participants select the target item.

The participants were asked to find and select the target item “Carpet” from the menu as quickly and accurately as possible. The target item was presented as a third item and located in the third group. To avoid the target position effect on the menu search time, the target item was located in the same position in both conditions (semantically organised menu and randomly organised menu), see Figure 4.1 (a) and Figure 4.1 (b).

4.2.5 Procedure

The study was conducted in a lab at the College of Computer and Information Science (CCIS) at King Saud University (KSU). The lab was quiet and free from any source of disturbance. The lab was equipped with 30 Dell desktops that run Windows OS. The two versions of the website were accessed through the Google Chrome web browser.

The study was conducted in two sessions on the same day. In each session, there were two groups. The first group was given a link to the first version of the website, which contained the semantically organised menu (the first condition). The second group was given a link to the second version of the website, which contained the randomly organised menu (the second condition). That means each participant was presented with one condition of menu organisations. The participants were organised into two groups by counterbalancing the two conditions between the participants. For example, the first sitting participant was assigned to the first group, and the next one was assigned to the second group, and so on.

The participants were introduced to the experiment by giving them brief information about the purpose of this study. Also, an information sheet and a consent form were given to the participant to read and sign, see Appendix B.2. Then, the demographic survey was given to the participants to be filled in.

After that, they were instructed to open the Google Chrome web browser and type the given link that was printed in the information sheet. After making sure that every participant opened the given link of the specific website according to the assigned group, all participants were instructed to start the experiment and select the target item, which was “Carpet”, as fast and accurate as possible. The participants were not informed about the organisation of the menu that they will search in.

To avoid making participants nervous by telling them about measuring their time on task, and to avoid making participants relaxed by not mentioning the time measurement, a good balance is to ask the participants to complete the task as quickly and accurately as possible without mentioning that time is recorded (Albert & Tullis, 2013).

After that, the collected data, which was saved in the databases of the website, was exported to excel sheets to be ready for the analysis process.

4.2.5 Pilot

The study was piloted by asking two colleagues to come to the lab that was reserved for conducting this experimental study. They were given brief information about the study. After that, each one was assigned to one condition. They opened the given link of the website and searched the displayed menu to find the target item. They did not have any comments.

4.2.6 Data Analysis

The collected data were stored in Excel sheets. These data were the participants’ performance data in the menu search task, such as task start time, task end time, search time, and selected item.

The boxplot was used to display the distribution of the search time data and to show the outliers in both menu conditions (semantically organised menu and randomly organised menu).

To test the relationship between menu organizations and outlying menu search performance, the Chi-Square test was considered. However, one assumption of this test was not met which is the expected values should be at least 5. Outliers are very low in a dataset, therefore, the expected values are less than 5.

Since this study is exploratory, the relationship between menu organizations and outlying menu search performance was tested by checking the number of outliers and their extreme level in each menu organization condition.

4.3 Results

The menu search time was measured in seconds. The average search time for each menu organisation condition was presented in Table 4.1.

Menu organisation condition	**Average Search Time**
Semantically organised menu	μ = 3.56 s, σ = 1.25
Randomly organised menu	μ = 5.75 s, σ = 5.71

Table 4.1: The average menu search time in each menu organisation condition.

The boxplots show outliers in both menu organisation conditions, see Figure 4.3. The total number of the outlying menu search performance cases was 3 cases, one case in the semantically organised menu (5.92 IQR) and two extreme cases in the randomly organised menu (6.88 IQR, 8.9 IQR).

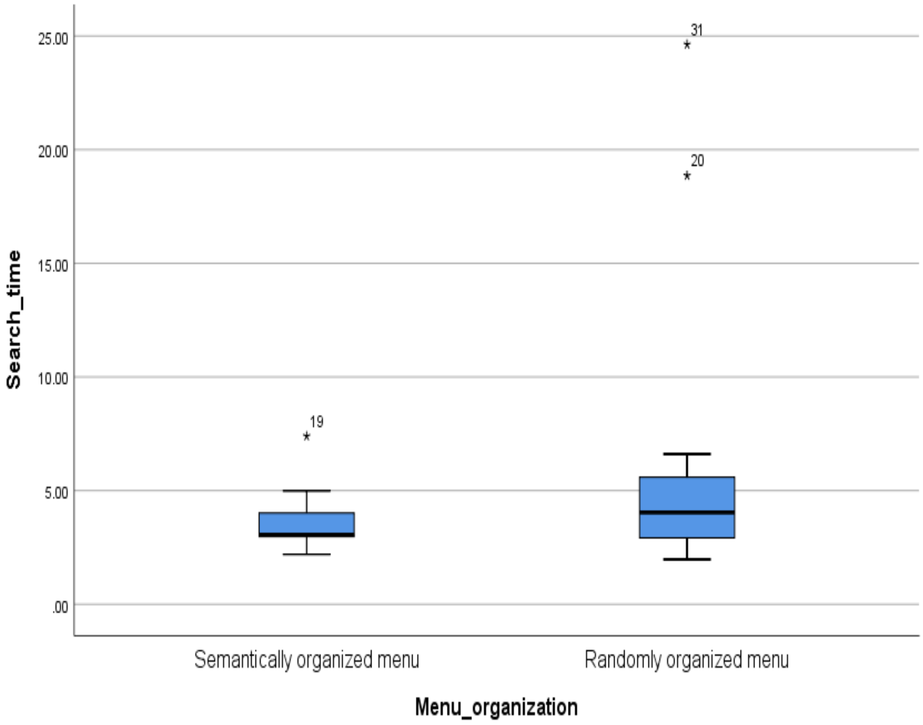


Figure 4.3: Boxplots of menu search time in the two menu organisation conditions. The outliers are represented by the asterisk at the top of the boxplots.

4.4 Discussion

The present study aimed to investigate whether the poor semantic organisation of menu items plays a role in outlying menu search performance. It hypothesised that more outliers are found in searching a poorly organised menu that visually groups semantically unrelated items.

This study found more outlying menu search performance cases in searching the poorly organised menu. Moreover, these cases were more extreme than the case found in searching the semantically organised menu. These results might suggest that poor semantic organisation of menu items, at least as perceived by some users, may play a role in outlying menu search performance. However, it could be that some confounding factors such as participants' variability in speed caused these outlying performance cases. It was difficult in this study to unconfound the effect of menu semantics and the effect of a slow user on performance time. Therefore, more studies using different methods are needed to check the role of the perceived menu semantic in outlying menu search performance.

The occurrence of outliers in this study that involved a simple menu search task confirmed the observation made by Schiller and Cairns (2008) that there is always one user who is extraordinarily slower compared to the other user.

Chapter 5 Collecting Semantic Similarity Ratings of Menu Items

5.1 Introduction

The finding of Study 2 (Chapter 4) might indicate that menu semantics may play a role in outlying menu search performance. However, more studies are needed to check the role of menu semantics in outlying performance using different methods. Therefore, a modelling approach was considered to help in understanding the role of the menu semantics in outlying menu search performance. Chen et al. (2015) menu search model was chosen. This model was presented in detail in Chapter 2 (section 2.6.1). As this model is based on ML, it needs to be trained on menu samples to learn menu search strategies. Constructing menu samples needs collecting semantic similarity ratings of menu items from human participants. Therefore, this chapter aimed at collecting the semantic similarity ratings of menu items from human participants.

One method for collecting the semantic similarity ratings of menu items is pairwise similarity ratings. In this method, participants are asked to rate the semantic similarity of each possible two items combination. Another method for collecting the semantic similarity ratings of menu items is card sorting. In this method, participants are asked to sort a set of items into groups of items that are similar in some way. Both methods help in identifying logical relationships between menu items (Bailly et al., 2016).

Chen et al. (2015) used the pairwise similarity ratings method to collect the semantic similarity ratings of menu items that were used to construct menu samples in their study. However, it was interesting to use both methods to collect the semantic similarity ratings of menu items and see whether they introduce different features on the data. This is important when populating a model based on this data. Therefore, it was decided to use both methods to collect the semantic similarity ratings of menu items. These collected semantic similarity ratings will be used afterwards to construct menu samples that are used in training the menu search model in Study 5 (Chapter 6).

This chapter presents two studies to collect the semantic similarity ratings of menu items: Study 3 and Study 4. Study 3 used pairwise similarity ratings to collect the semantic similarity ratings from participants. Study 4 used card sorting to collect the semantic similarity ratings from participants. Additionally, this chapter utilised the results of these two studies and made a comparison between the two methods: pairwise similarity ratings and card sorting.

5.2 Study 3: Collecting the Semantic Similarity Ratings of Menu Items Using Pairwise Similarity Ratings Method

The pairwise similarity ratings method is thought of as the gold standard method for producing a similarity dataset (Lantz et al., 2019). It was used for collecting the semantic similarity ratings of menu items in previous menu search modelling studies such as Bailly et al. (2014) and Chen et al. (2015). Therefore, this study used this method to collect the semantic similarity ratings of menu items.

5.2.1 Method

Design

This study aimed at collecting the semantic similarity ratings of menu items using the pairwise similarity ratings method. Although this method was used by previous menu search modelling studies such as Bailly et al. (2014) and Chen et al. (2015), no details were found in these studies regarding how this method was applied and how the collected data from this method were processed. Therefore, this study was designed based on common research methods.

A survey was used as a way to present item pairs and rating scales. The survey was the most appropriate method for collecting the users' semantic similarity ratings. This is because the survey allows asking closed questions that use rating scales as a response format. Rating scales are good for making respondents do assessments of things (Preece et al., 2015).

A web-based survey was adopted instead of a paper-based survey because the web-based survey eliminates the need for manual data entry, which is time-consuming and prone to data entry errors (Lazar et al., 2017).

The web-based survey in this study consisted of a list of closed questions. Each question presented a pair of items and a five-point rating scale. There were five ordered rating categories responses that represent degrees of closeness in meaning ("not at all close", "only a little close", "somewhat close", "quite close", and "very close"). The respondents were asked to note their ratings of to what extent the two items in the pair are close in meaning. A five-point rating scale was used because five is a medium-sized range that seems suitable when assessing semantic similarity. It was used by Schiller and Cairns (2008) to collect the semantic relevance assessment of menu items for relevance to the task.

The questions' order was randomised to avoid bias that might be caused by respondent fatigue. Respondent fatigue might occur during the later questions due to the tiredness of the survey task, and that could affect the quality of the provided data (Lavrakas, 2008). Randomising the order of questions in the survey is frequently used to prevent the bias introduced by respondent fatigue (Hillmer, 2020).

Participants

This study was conducted at the CCIS at the Female Campus at KSU. Twenty-nine participants took part in this study. The participants' age range was between 20 and 45, with a mean age of 26.3 (SD = 7.6). Two-thirds of the participants were students in the CCIS at KSU, while the rest of the participants were administrative staff or academic staff at the same College. They were all female as KSU adopts a single-gender education. They all were native Arabic speakers. The participants were not the same participants who participated in Study 2 (Chapter 4). All participants were asked in person if they are willing to participate in the study.

Task and Materials

The participants were asked to rate 120 pairs of menu items according to what they thought about how close in meaning the two items in the pair are. The primary material in this study was an online survey consisting of 120 rating questions that ask to rate the semantic similarity of 120 pairs of menu items, see Appendix C.2. The items used in this survey were the same 16 menu items used in Study 2 (Chapter 4). These items are listed in Table 5.1, along with their Arabic translation.

Category	Item in English	Item in Arabic
Appliance	Computer	جهاز الحاسوب
	Dishwasher	غسالة صحون
	Refrigerator	ثلاجة
	Television	تلفاز
Jewellery	Bracelet	سوار
	Ring	خاتم
	Cufflink	كبك رجالي
	Crown	تاج
Furniture	Sofa	اريكة
	Table	طاولة
	Carpet	سجادة
	Lamp	مصباح
Room in house	Basement	قبو
	Bedroom	غرفة نوم
	Balcony	شرفة
	Kitchen	مطبخ

Table 5.1: The 16 items used in the pairwise similarity ratings survey.

These sixteen selected items were translated into the Arabic language and mixed to form 120 different pairs of menu items. Examples of pairs used in the survey are: (Computer - Ring), (Sofa - Balcony), and (Bracelet - Crown).

The survey was built by using Qualtrics which is an online survey tool that offers a lot of features to build and distribute surveys as well as analyse responses. Qualtrics was selected because it is reliable and has many advanced features. Additionally, the University of York has a license to allow students and staff to use this survey tool.

Procedure

This study was conducted in the researcher's office at CCIS at KSU. The study was conducted in several sessions. Participants came in groups of 3 to 5 participants. Once the participants arrived at the office, they first were welcomed and given a brief introduction about the purpose of this study and what they would be requested to do. After that, the information sheet and informed consent were given to the participant to read and sign, see Appendix C.1. Then, the participants were asked to answer the demographic questionnaire, which included questions about personal details such as age and educational level. After that, the participants were given a link to the online survey and asked to open this link and start answering the questions. These questions were about rating the semantic similarity of 120 pairs of items. After completing the survey, the participants were thanked for their participation. Overall, the average time taken by the participants to answer the survey questions was 14 m 27 s.

Pilot

It is important to pilot a survey study (also called pretesting the survey) to make sure that questions are understandable. Two aspects need to be tested during the pilot study: the survey questions and the survey interface design (Lazar et al., 2017).

Dillman (2000, as cited in Lazar et al., 2017) suggested three steps procedure for pretesting a survey and stated that it is rarely applied thoroughly. These three steps are as follows: asking experienced colleagues to review the survey, asking potential respondents to assess the clarity and motivation of the questions, and conducting a pilot study to test the survey tool and procedures.

To pilot this study, first, the survey interface design was evaluated by two knowledgeable colleagues. They suggested changing the font colour of the questions to differentiate them from the answers. Their comment was considered, and the survey was edited to reflect this suggestion.

Second, two potential respondents were asked to complete the entire survey to ensure that the questions are unambiguous and to test the usability of the interface. It was noted that one of the respondents asked for examples of two items that are close in meaning and another two items that are not close in meaning. Therefore, this issue was considered, and two examples were added to the information sheet to be read and understood before answering the survey questions. These examples were (Mountain - Valley) and (Uncle - Brother) as examples of items that are close in meaning and (Apple - Diamond) and (Cat - Chair) as examples of items that are not close in meaning. The items in these examples were not from the items list used in the study's survey.

Data Analysis

For each pair of items in the survey, the rating responses were used to calculate the average semantic similarity between the items in this pair. I used a weighted average because I considered some rating responses to be more important than others and should be contributed more to the final average similarity. Therefore, a weight was given for each rating response as follows (very close = 100, quite close = 75, somewhat close = 50, only a little close = 25 and not at all close = 0). Then, I multiplied the number of individuals selecting each rating with the corresponding rating weight. After that, I added the results of those calculations together and divided the result by the number of responses. I combined these calculations in one formula as follows:

Average similarity between items in the pair = N(very close) * 100 + N(quite close) * 75 + N(somewhat close) * 50 + N(only a little close) * 25 + N(not at all close) * 0 / Number of participants.

Where:

N(very close) = number of participants who selected "very close" option.

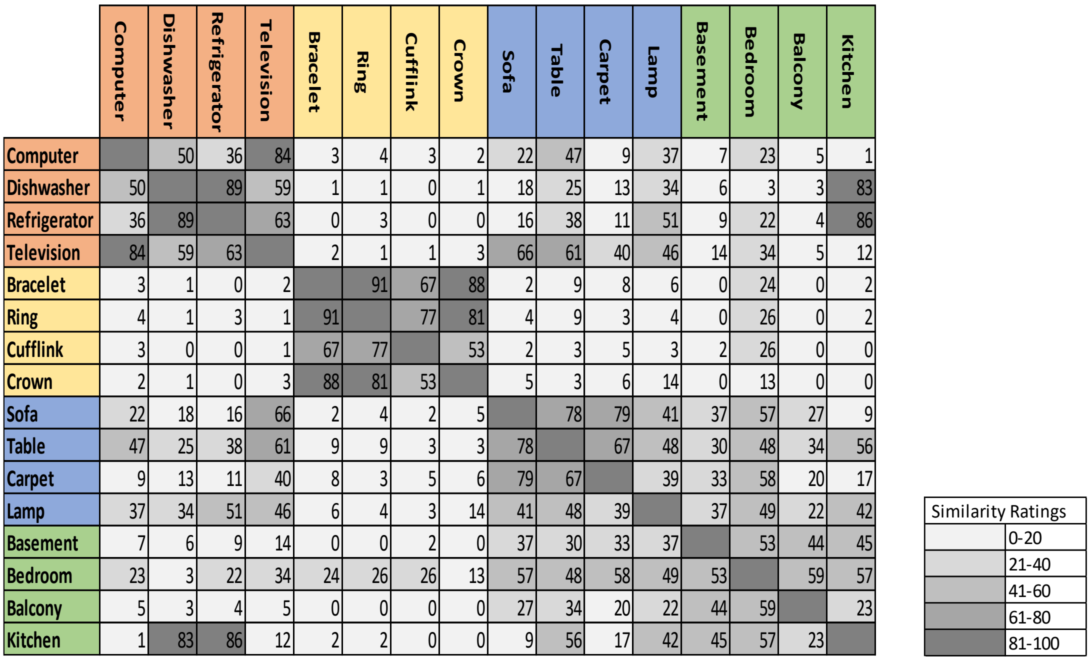
N(quite close) = number of participants who selected "quite close" option.

N(somewhat close) = number of participants who selected "somewhat close" option.

N(only a little close) = number of participants who selected "only a little close" option.

N(not at all close) = number of participants who selected "not at all close" option.

After calculating the average similarity for the 120 pairs of menu items, the 16 by 16 similarity matrix was created to represent the semantic similarity between the 16 menu items based on the previous calculations of the average similarity for each pair of items in the survey, see Figure 5.1.



	Computer	Dishwasher	Refrigerator	Television	Bracelet	Ring	Cufflink	Crown	Sofa	Table	Carpet	Lamp	Basement	Bedroom	Balcony	Kitchen
Computer		50	36	84	3	4	3	2	22	47	9	37	7	23	5	1
Dishwasher	50		89	59	1	1	0	1	18	25	13	34	6	3	3	83
Refrigerator	36	89		63	0	3	0	0	16	38	11	51	9	22	4	86
Television	84	59	63		2	1	1	3	66	61	40	46	14	34	5	12
Bracelet	3	1	0	2		91	67	88	2	9	8	6	0	24	0	2
Ring	4	1	3	1	91		77	81	4	9	3	4	0	26	0	2
Cufflink	3	0	0	1	67	77		53	2	3	5	3	2	26	0	0
Crown	2	1	0	3	88	81	53		5	3	6	14	0	13	0	0
Sofa	22	18	16	66	2	4	2	5		78	79	41	37	57	27	9
Table	47	25	38	61	9	9	3	3	78		67	48	30	48	34	56
Carpet	9	13	11	40	8	3	5	6	79	67		39	33	58	20	17
Lamp	37	34	51	46	6	4	3	14	41	48	39		37	49	22	42
Basement	7	6	9	14	0	0	2	0	37	30	33	37		53	44	45
Bedroom	23	3	22	34	24	26	26	13	57	48	58	49	53		59	57
Balcony	5	3	4	5	0	0	0	0	27	34	20	22	44	59		23
Kitchen	1	83	86	12	2	2	0	0	9	56	17	42	45	57	23	

Similarity Ratings	
░	0-20
▒	21-40
▒	41-60
▓	61-80
█	81-100

Figure 5.1: The 16-by-16 semantic similarity matrix generated by the pairwise similarity ratings method.

The semantic similarity matrix was then analysed using statistical clustering techniques such as hierarchical clustering and multidimensional scaling (MDS). These techniques were used to identify the underlying clusters (semantic groups of menu items) according to the participants' semantic similarity ratings of menu items.

Clustering functions are provided by common statistical analysis tools such as R and SPSS. Before applying any clustering function in R, a dissimilarity, or distance matrix, should be created (Kassambara, 2017). To create a distance matrix using the R, the dist() function was used to compute the distance matrix based on the input similarity matrix. After that, to create a hierarchical clustering dendrogram using R, the hclust() function was used to generate clusters based on the distance matrix. To create an MDS plot using R, the mds() function was applied to the distance matrix to visualise the distance between all pairs of items in two dimensions.

5.2.2 Results

- **The resulting semantic groups**

The hierarchical clustering dendrogram showed three clusters (semantic groups): 1) jewellery, 2) furniture and room in house, 3) kitchen, and appliance, see Figure 5.2.

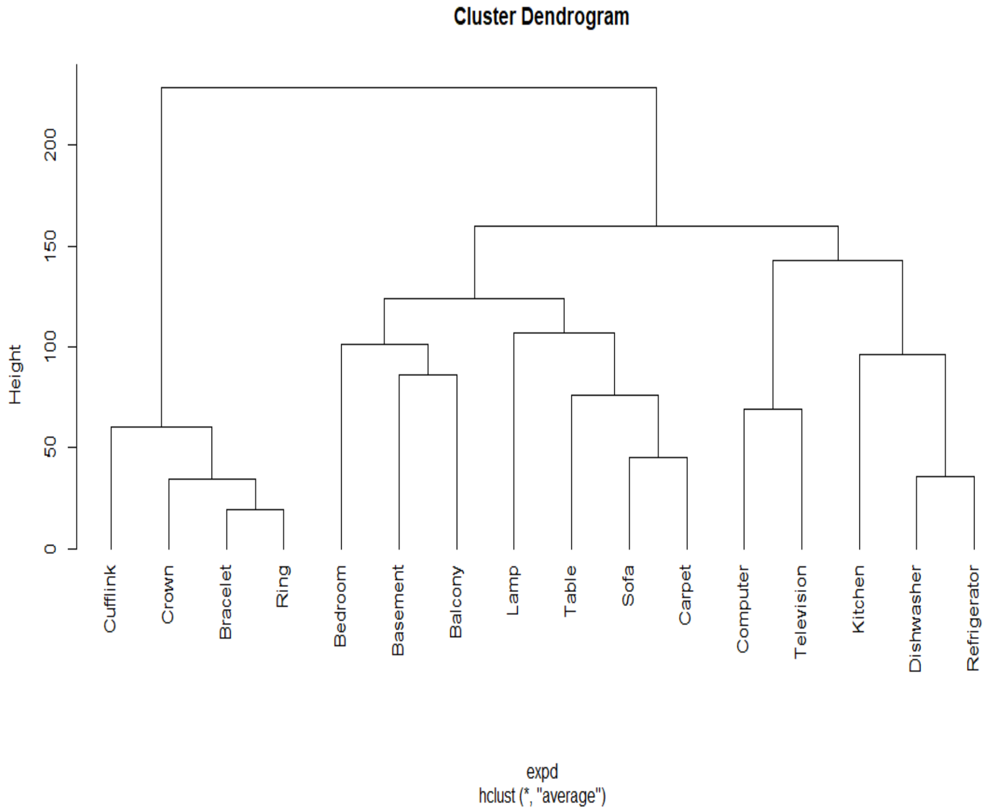


Figure 5.2: Three clusters (semantic groups) are shown in hierarchical cluster dendrogram, 1) jewellery on the left, 2) furniture and room in house on the middle, and 3) kitchen and appliance on the right.

The MDS plot also showed the same three clusters (semantic groups): 1) jewellery in the right side of the MDS plot, 2) furniture and room in the house in the lower left of the MDS plot, and 3) kitchen and appliance in the upper left of the MDS plot, see Figure 5.3. The two clusters on the left side of the MDS plot were blurred. This might be because the items in these two clusters (furniture and room in the house, kitchen and appliance) seemed related to each other as they all related to the house although they belonged to different original natural categories.

The resulting semantic groups were different compared to the original natural categories (Appliance, Jewellery, Furniture and Room in house).

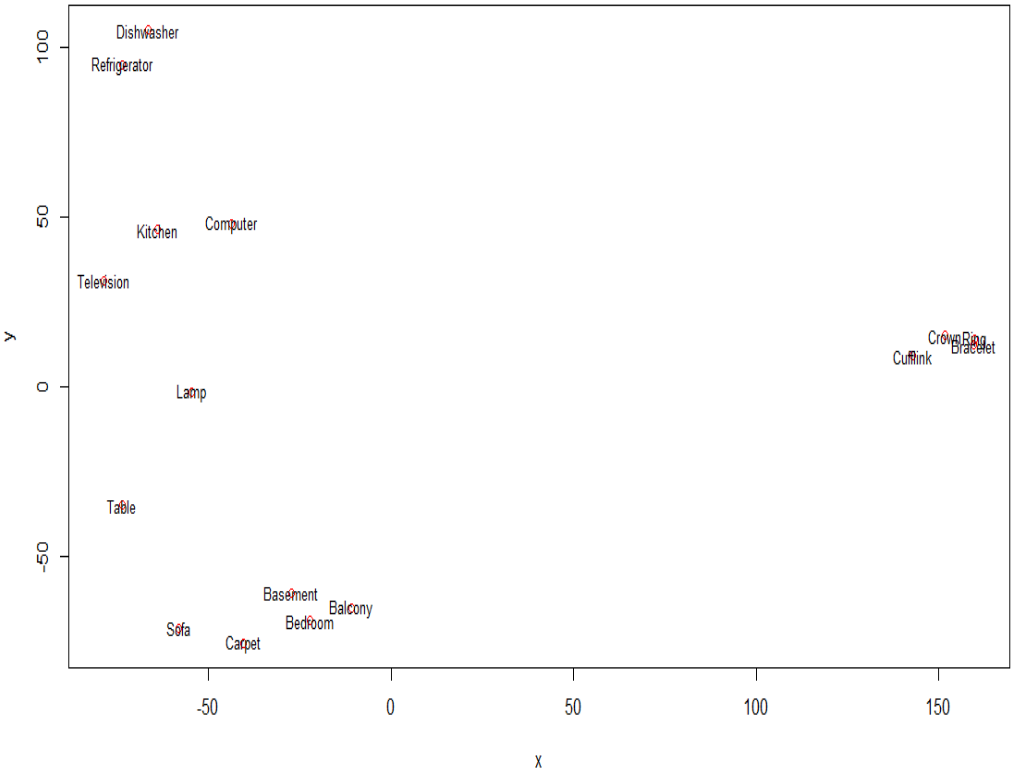


Figure 5.3: Three clusters (semantic groups) are shown in multidimensional scaling (MDS) plot, 1) jewellery on the right side, 2) furniture and room in the house on the lower left, and 3) kitchen and appliance in the upper left.

- **The individual's ratings**

Some participants rated the semantic similarity of menu items differently; for example, the average semantic similarity between computer and bracelet was 2.19, see Figure 5.1, these two items belong to different original natural categories and seem semantically dissimilar, but two participants rated them as very close, see Figure 5.4.

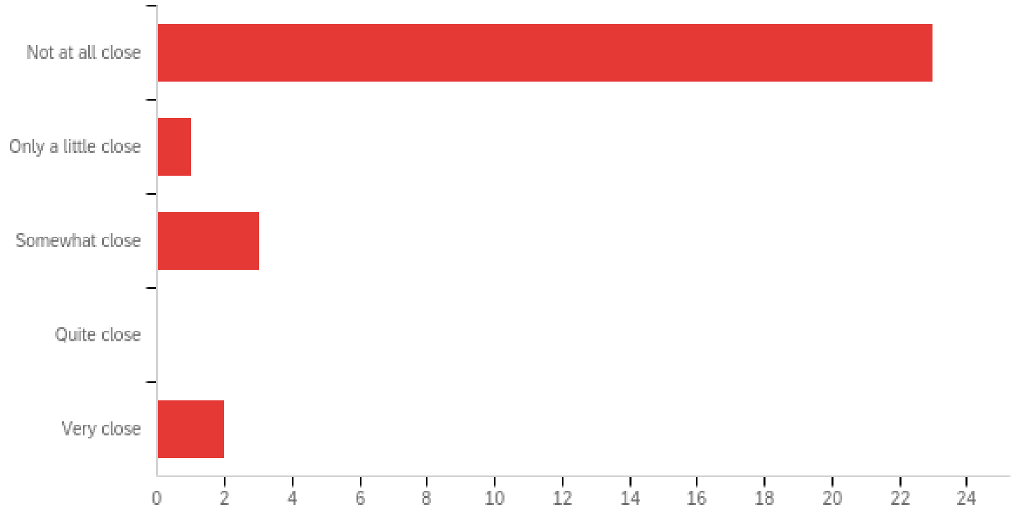


Figure 5.4: Computer and Bracelet semantic similarity ratings.

Another example was the average similarity between sofa and table was 77.59, see Figure 5.1. Although these two items belong to the same original natural category and seem semantically similar, there was one participant who rated them as not at all close, see Figure 5.5.

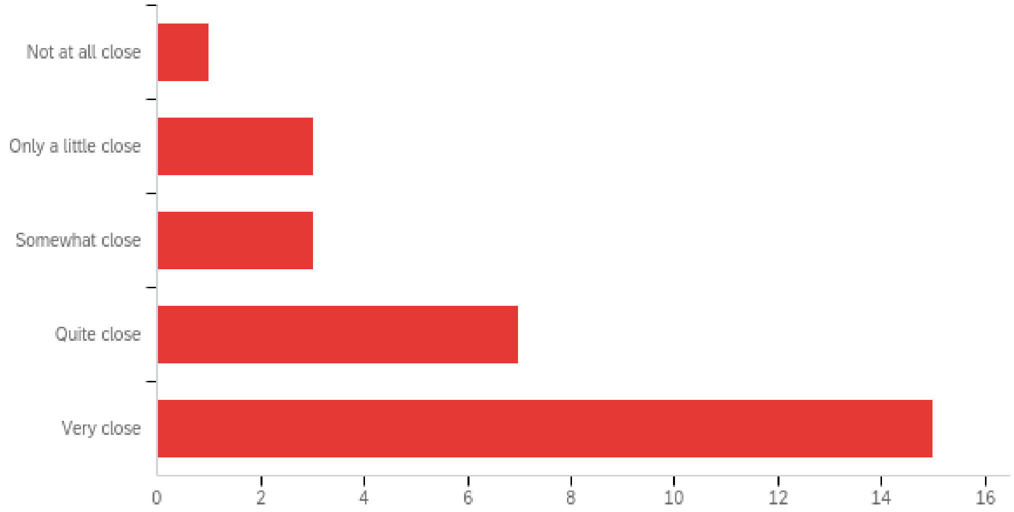


Figure 5.5: Sofa and Table semantic similarity ratings.

Also, it was noticed that the average semantic similarity between refrigerator and kitchen was 86.2, and most participants rated them as very close or quite close, see Figure 5.6, although they do not belong to the same original natural category. This might be because these two items are related to each other.

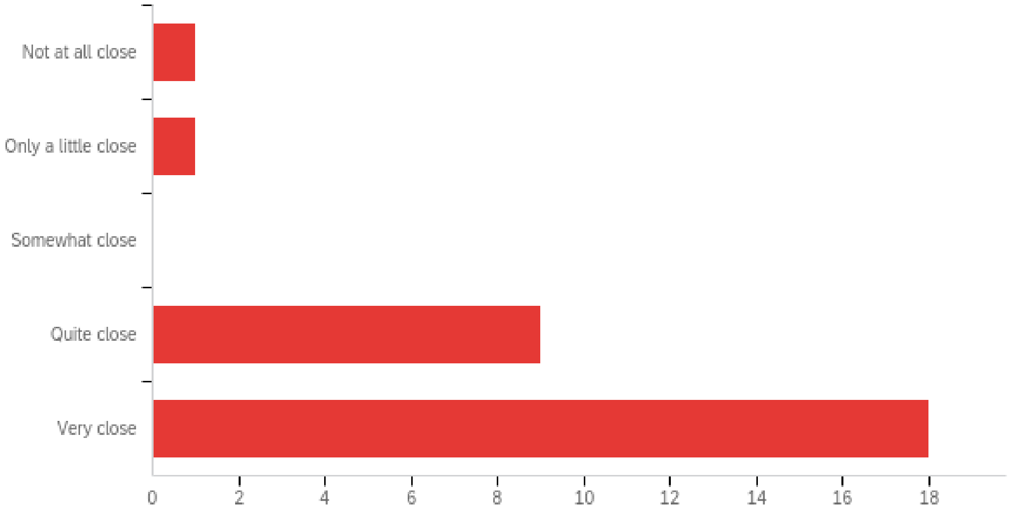


Figure 5.6: Refrigerator and Kitchen semantic similarity ratings.

Also, there were some pairs of items that were rated differently by the participants, such as Television and Carpet, see Figure 5.7. Almost 51.72% of the participants rated them to be not at all close or only a little close, while 27.58% of participants rated them as very close or quite close.

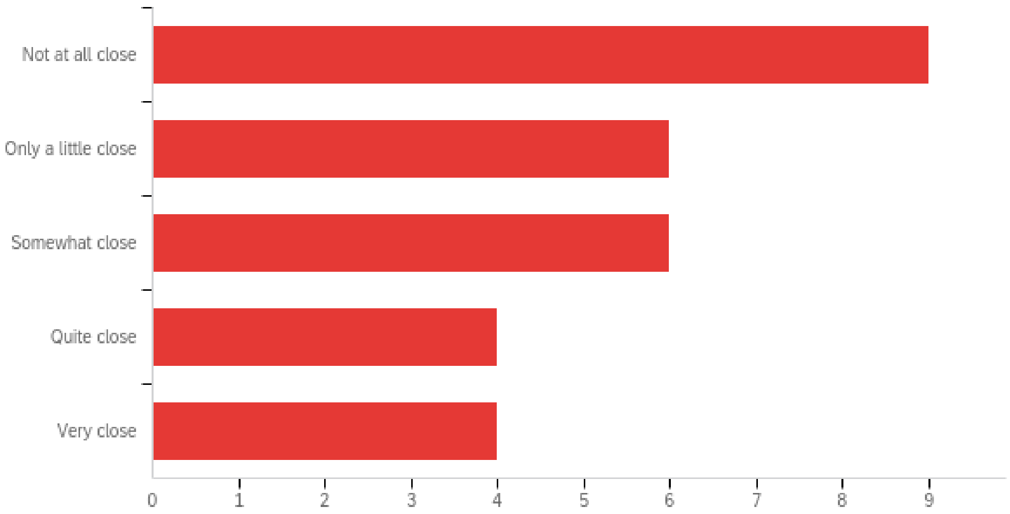


Figure 5.7: Television and Carpet semantic similarity ratings.

5.2.3 Discussion

This study used the pairwise similarity ratings method to collect the semantic similarity ratings of menu items. Participants were given a survey consisting of 120 pairs of menu items and asked to rate each pair according to how close in meaning the two items in the pair are. The participant's ratings were processed to produce the semantic similarity dataset that was analysed using clustering techniques to find out the resulting semantic groups.

The resulting semantic groups were different compared with the original natural categories that menu items were derived from. This might be because some pairs of items such as kitchen and refrigerator were rated by many participants as semantically similar, although they belong to different original natural categories (Appliance and Room in house). These items seem related to each other, but they are not semantically similar. Resnik (1999) makes a distinction between "similarity" and "relatedness" and gives an example: cars and gasoline seemed more tightly related than cars and bicycles, but the last pair is more similar. Therefore, it could be that some participants mixed semantic similarity with semantic relatedness, although I clarified what I meant by the semantic similarity in the information sheet of the study by giving examples of pairs of items that are semantically similar and pairs of items that are semantically dissimilar.

Additionally, the results of this study indicate that there are individual differences in perceiving the semantic similarity of menu items.

5.3 Study 4: Collecting the Semantic Similarity Ratings of Menu Items Using Card Sorting method

Card sorting is a common method in HCI research and practice. It involves requesting participants to sort a set of items into groups of items that are similar in some way. Card sorting assesses the extent of the perceived semantic similarity within a group of items (Schmettow & Sommer, 2016). Therefore, this study used the card sorting method to collect the semantic similarity ratings of menu items.

5.3.1 Method

Design

This study used open card sorting to collect the semantic similarity ratings of menu items. An open card sorting was adopted in this study because of the need to give participants the freedom to form different groups of items that reflect their assessments of the semantic similarity of these menu items.

Online card sorting was adopted in this study. An online card sorting was adopted because of the need to save time and effort for both the researcher and participants. This is because software-based card sorting, such as online card sorting, is a one-step process as no need to process and analyse the results using another tool (Spencer, 2009).

Individual card sorting was applied, so each participant sorted the cards separately. This allowed collection of diverse responses. These diverse responses are important to understand how people are different in perceiving the semantic similarity of menu items.

Participants

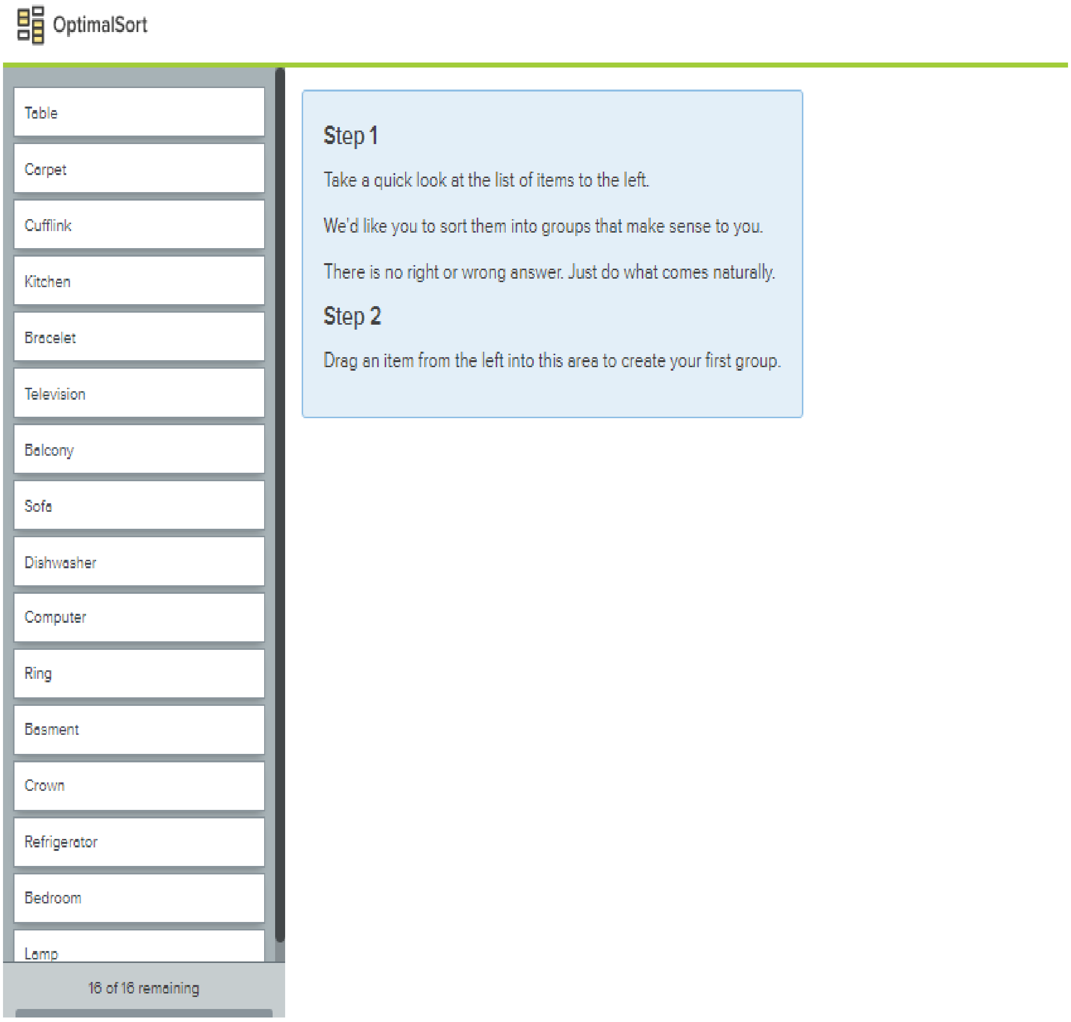
This study was conducted at the CCIS at the Female Campus at KSU. Twenty-six participants took part in this study. The participants' ages range between 20 and 30, with a mean age of 21.15 years (SD = 1.9). They were all female as KSU adopts a single-gender education. They were all native Arabic speakers. The participants were not the same participants who participated in Study 2 (Chapter 4) and Study 3. All participants were asked in person if they were willing to participate in the study.

Task and Materials

The participants were asked to group the presented cards into groups according to what they thought about how close in meaning these items in these cards are.

The items that were used in this study were the same items used in Study 3, see Table 5.1. The sixteen items were translated into the Arabic language and mixed to form the cards that were sorted by the participants, see Appendix C.4.

The primary material in this study was online card sorting. The OptimalSort tool was used to create the online card sorting, see Figure 5.8. The OptimalSort is an online card sorting tool that is offered by the Optimal Workshop user research platform.



OptimalSort

Table
Carpet
Cufflink
Kitchen
Bracelet
Television
Balcony
Sofa
Dishwasher
Computer
Ring
Basment
Crown
Refrigerator
Bedroom
Lamp

16 of 16 remaining

Step 1

Take a quick look at the list of items to the left.

We'd like you to sort them into groups that make sense to you.

There is no right or wrong answer. Just do what comes naturally.

Step 2

Drag an item from the left into this area to create your first group.

Figure 5.8: The online card sorting task created by the OptimalSort tool.

Procedure

The study was conducted in a quiet lab. This lab was equipped with 30 Dell computers that run Windows OS. This study was conducted in one session. When participants arrived at the lab, they were first welcomed and asked to find a seat in the lab and open the computers to be ready. Then, the participants were given a brief introduction of the purpose of this study and what they would be requested to do. After that, the information sheet and informed consent were given to the participant to read and sign, see Appendix C.2.

After that, the participants were asked to open the online card sorting link and start the sorting task. After finishing the task, the participants were thanked for their participation. Overall, the average time taken by the participants to sort the cards was 5 m 21 s.

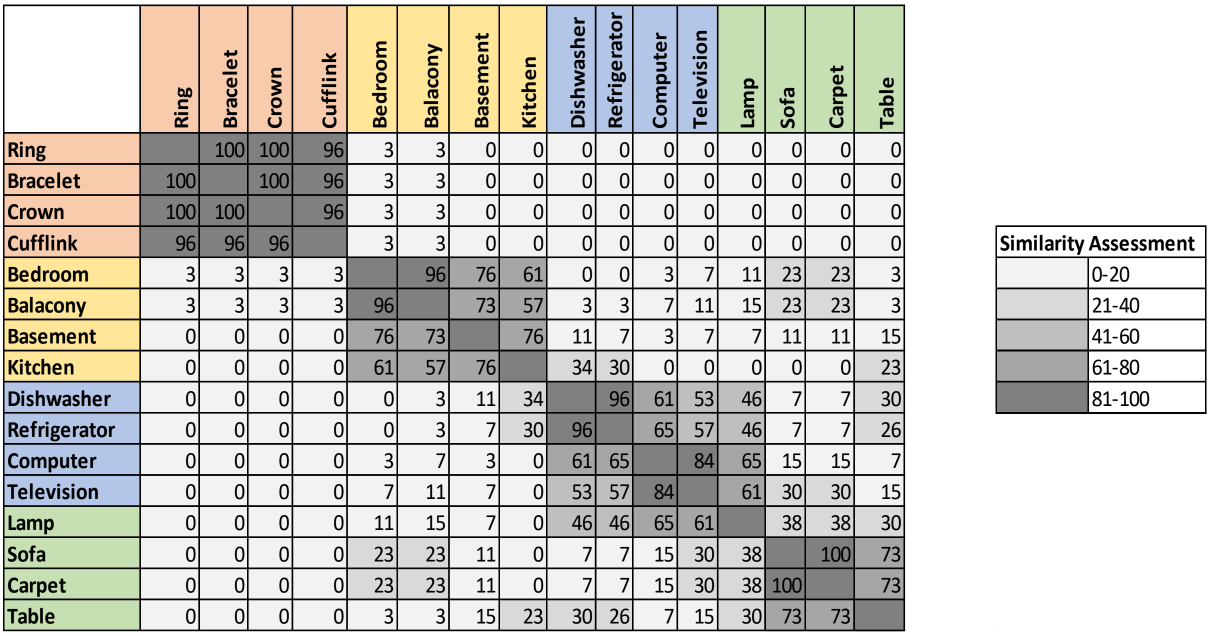
Pilot

Before conducting this study, it was important to test the material and the procedure. Two colleagues participated in the pilot study. They carried out the task without any problem, and they had no comments.

Data Analysis

The 16-by-16 similarity matrix was created automatically by the Optimalsort tool. The similarity values in the similarity matrix represented the percentage of participants who put these two items in the same group, see Figure 5.9.

The same clustering techniques used in Study 3 were also applied in this study to analyse the similarity matrix.



	Ring	Bracelet	Crown	Cufflink	Bedroom	Balacony	Basement	Kitchen	Dishwasher	Refrigerator	Computer	Television	Lamp	Sofa	Carpet	Table
Ring		100	100	96	3	3	0	0	0	0	0	0	0	0	0	0
Bracelet	100		100	96	3	3	0	0	0	0	0	0	0	0	0	0
Crown	100	100		96	3	3	0	0	0	0	0	0	0	0	0	0
Cufflink	96	96	96		3	3	0	0	0	0	0	0	0	0	0	0
Bedroom	3	3	3	3		96	76	61	0	0	3	7	11	23	23	3
Balacony	3	3	3	3	96		73	57	3	3	7	11	15	23	23	3
Basement	0	0	0	0	76	73		76	11	7	3	7	7	11	11	15
Kitchen	0	0	0	0	61	57	76		34	30	0	0	0	0	0	23
Dishwasher	0	0	0	0	0	3	11	34		96	61	53	46	7	7	30
Refrigerator	0	0	0	0	0	3	7	30	96		65	57	46	7	7	26
Computer	0	0	0	0	3	7	3	0	61	65		84	65	15	15	7
Television	0	0	0	0	7	11	7	0	53	57	84		61	30	30	15
Lamp	0	0	0	0	11	15	7	0	46	46	65	61		38	38	30
Sofa	0	0	0	0	23	23	11	0	7	7	15	30	38		100	73
Carpet	0	0	0	0	23	23	11	0	7	7	15	30	38	100		73
Table	0	0	0	0	3	3	15	23	30	26	7	15	30	73	73	

Similarity Assessment	
□	0-20
░	21-40
▒	41-60
▓	61-80
█	81-100

Figure 5.9: The 16-by-16 similarity Matrix generated by card sorting.

5.3.2 Results

- **The resulting semantic groups**

Four clusters (semantic groups) were found in the hierarchical cluster dendrogram and the MDS plot; see Figure 5.10 and Figure 5.11. These four semantic groups are as follows: jewellery, room in a house, furniture, appliance and lamp. These semantic groups were very similar to the original natural categories that items derived from (appliance, jewellery, furniture, and room in house), except the lamp item was grouped with appliance category items. This might be because the lamp item might be related to the appliance to some extent.

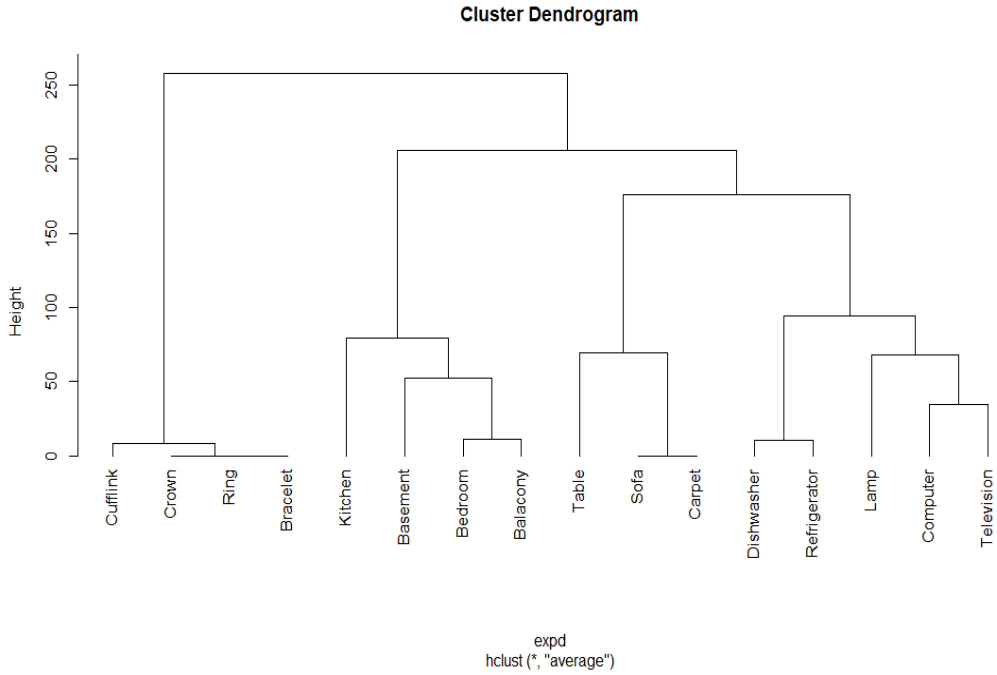


Figure 5. 10: Four clusters (semantic groups) are shown in the hierarchal cluster dendrogram, 1) jewellery on the left, 2) room in house on the middle-left, 3) furniture on the middle-right and 4) appliance and lamp on the right.

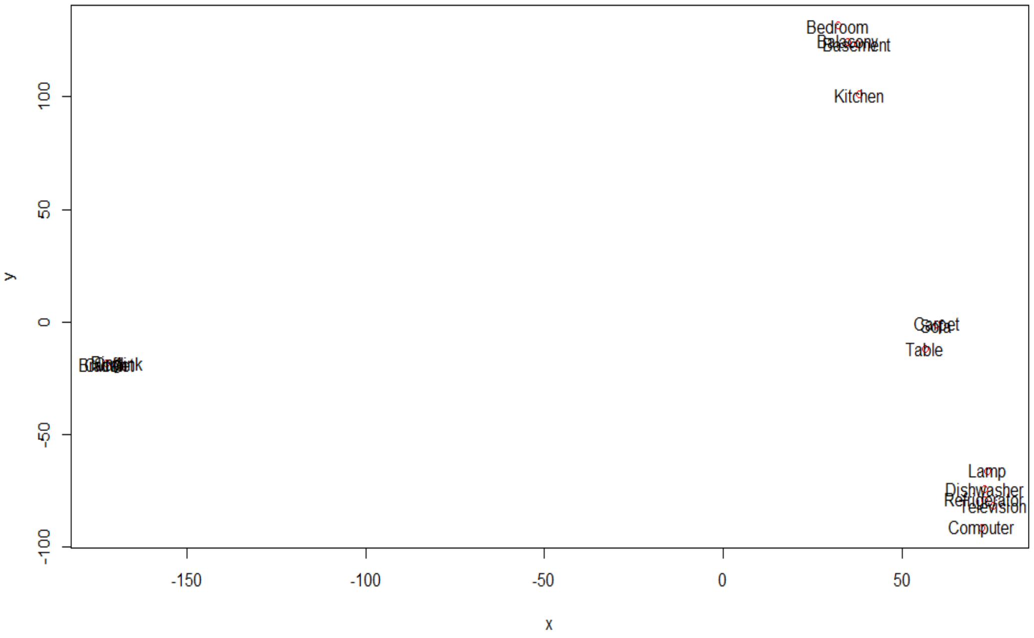


Figure 5.11: Four clusters (semantic groups) are shown in multidimensional scaling (MDS), 1) jewellery on the left side, 2) rooms in the house on the upper right, 3) furniture on the middle right and 4) appliances and lamp on the lower right.

- **The Individual’s sorting**

Some participants were different in sorting the items; for example, there were two participants (7%) who put computer and balcony in the same group. Also, there were two participants (7%) who put and refrigerator and carpet in the same group. Also, there was one participant (3%) who grouped ring and bedroom, see Figure 5.9.

5.3.3 Discussion

This study used the card sorting method to collect the semantic similarity ratings of menu items. Participants were asked to sort 16 menu items into groups according to what they thought about how close in meaning these items are. The participants’ sorting was used to produce the similarity matrix that was analysed using clustering techniques to find out the resulting semantic groups.

The resulting semantic groups were very similar to the original natural categories. This might be because that card sorting presents all items together, so this gives a context that helps in perceiving the semantic similarity of menu items.

The results of this study showed that some participants sorted the items differently. This indicates that there are individual differences in perceiving the semantic similarity of menu items.

By comparing card sorting results with pairwise similarity rating results, the card sorting and pairwise similarity ratings generated different semantic groups. When using pairwise similarity ratings, there were three resulting semantic groups (1- appliance and kitchen, 2- jewellery, 3- furniture and room in the house). They were different compared with the four original natural categories (1- appliances, 2- jewellery, 3-furniture, 4- rooms in the house). When using card sorting, there were four resulting semantic groups (1-jewellery, 2-rooms in a house, 3-furniture and 4- appliances and lamp). They were very similar to the original natural categories compared with the resulting semantic groups that were found when using pairwise similarity ratings. This indicates that the card sorting method produces more meaningful semantic groups compared with the pairwise similarity ratings method. This might be because that card sorting allows presenting the whole set of items, so participants can organise and reorganise the items while performing the sorting task until reaching the optimal organisation. Additionally, full exposure to the entire set of items through the assessment process offers a grounding context (O'Shea et al., 2010).

The card sorting and pairwise similarity ratings yield different results, as evidenced statistically by the moderate correlation between these two methods, r = 0.641, r^2 = 0.411, see Figure 5.12. This confirms the findings of Dwyer (2003) and Lantz et al. (2019), who found that there was a moderate correlation between a card sorting similarity matrix and a pairwise similarity ratings matrix. It should be clarified that some assumptions of the applied correlation test (Pearson correlation test) might not be met such as the normality of variables. However, the r value was used here qualitatively rather than drawing a strong conclusion.

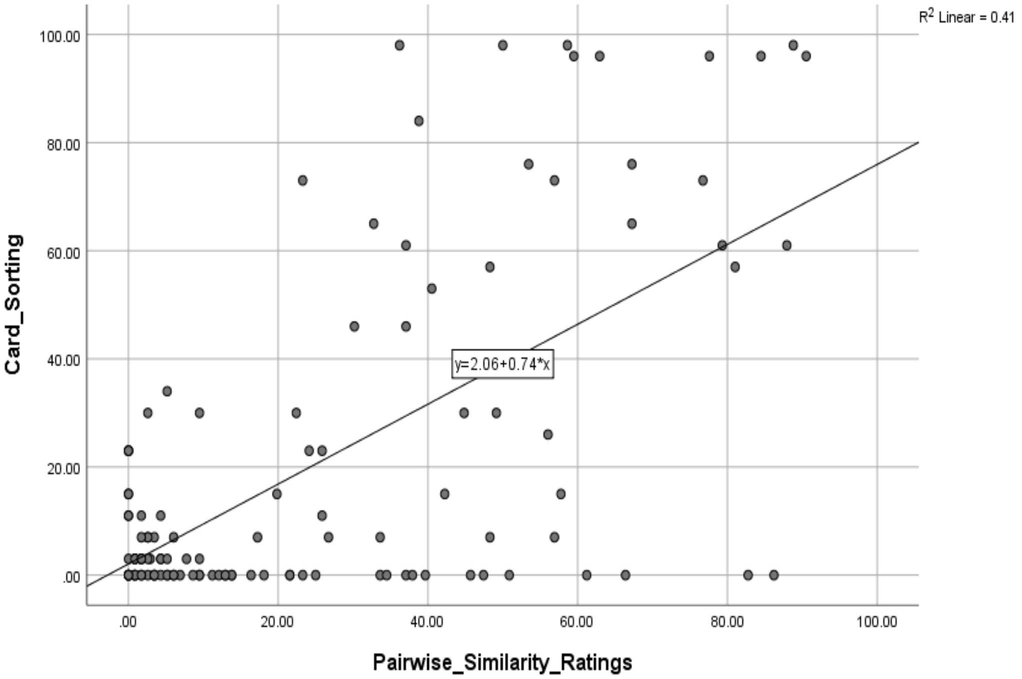


Figure 5.12: The scatterplot of card sorting data and pairwise similarity ratings data.

When using card sorting to collect the semantic similarity ratings of menu items, the participants spent less time completing the rating task compared with using a pairwise similarity rating. The time spent when using card sorting was about one-third of the time spent when using pairwise similarity ratings. This indicates that using card sorting reduced the time required to complete the rating task. This might be because card sorting allows participants to make decisions about the whole set of items at the same time. These concurrent decisions eliminate the multiple pairwise individual assessments that are required in pairwise similarity ratings (Lantz et al., 2019). The multiple pairwise assessments consume more time and demand more cognitive effort. This is considered a primary disadvantage of the pairwise similarity rating method as this might affect the quality of the collected data, as stated by Lantz et al. (2019).

5.4 Discussion

This chapter aimed to collect the semantic similarity ratings of menu items. The semantic similarity ratings are required to construct menu samples that are needed to train the menu search model in the next chapter (Chapter 6). Two studies were conducted to collect the semantic similarity ratings using two different methods. Study 3 used the pairwise similarity ratings method to collect the semantic similarity ratings of menu items, and Study 4 used the card sorting method to collect the semantic similarity ratings of menu items. Additionally, this chapter utilised the results of the two studies and made a comparison between the two methods: pairwise similarity ratings and card sorting.

By comparing these two methods, it seems that the card sorting method is better when it comes to eliciting users' perception of the menu structure. It produces meaningful semantic groups. Additionally, it is much faster. These advantageous features might justify why this method is commonly used in designing menu structures. However, when it comes to modelling user performance in menu search, the pairwise similarity ratings method seems more appropriate to collect the semantic similarity ratings of menu items that are used to implement the model's semantic similarity function and to construct menu samples. This is because, in this method, participants assess the semantic similarity of two items at a time, and that reflects what happens during menu search when users assess the semantic similarity of the fixated menu items and the target at a time. This might justify why this method was used to collect the semantic similarity ratings of menu items in the menu search modelling study by Chen et al. (2015).

The collected semantic similarity ratings in both studies showed that some participants rated the semantic similarity of menu items very differently. Therefore, it would be interesting to investigate whether people who rate the semantic similarity of menu items very differently perform poorly in menu search.

Chapter 6

Study 5: Investigating the Role of Menu Semantics in Outlying Menu Search Performance Using Modelling Approach

6.1 Introduction

A modelling approach was used in this research to help understand the role of menu semantics in outlying menu search performance. Chen et al. (2015) menu search model was chosen. This model was presented in detail in Chapter 2 (section 2.6.1). As this model is based on ML, it needs to be trained in a menu search environment to learn menu search strategies. Once the model learns the menu search strategies through the training, it can be used to predict the menu search performance.

This study aimed to test whether the adopted menu search model can predict outlying menu search performance due to menu semantics. Therefore, the model code was obtained from the author of this model (Dr Xiuli Chen). The model was implemented using the MATLAB platform. The model code was read to understand the main functions. After that, the model was trained and tested using the collected data in the previous chapter. First, the model was trained and tested using the semantic similarity data collected by the pairwise similarity ratings method. Then, the model was trained and tested using the semantic similarity data collected by the card sorting method.

6.2 Training and Testing the Model Using Data Collected by the Pairwise Similarity Ratings

In Study 3 (Chapter 4), the semantic similarity ratings of menu items were collected using the pairwise similarity ratings method. The collected semantic similarity ratings were prepared to be used in training and testing the model. The training and testing procedures are detailed in the following section.

6.2.1 Method

Procedure

The computer that was used to train and test the model was an ASUS desktop. This computer has an Intel(R) Core™ i7 processor (1.80 GHz) and 8.00 GB RAM. It runs a Windows 10 OS.

To train the model, I followed the same training procedures that were done by Chen et al. (2015). First, a menu search environment should be created to train the model. The menu search environment should be represented by the ecological distributions of menu length, item length and semantic group size. Also, by the ecological distribution of semantic relevance of menu items.

In this study, to create the menu search environment, the menu used in Study 2 (Chapter 4) was used. This menu was used to determine the ecological distribution of menu item length, see Figure 6.1, right panel. Also, it was used to determine the ecological distribution of the semantic similarity ratings of menu items, see Figure 6.1, left panel. These semantic similarity ratings were collected from human participants in Study 3 (Chapter 5) using the pairwise similarity ratings method.

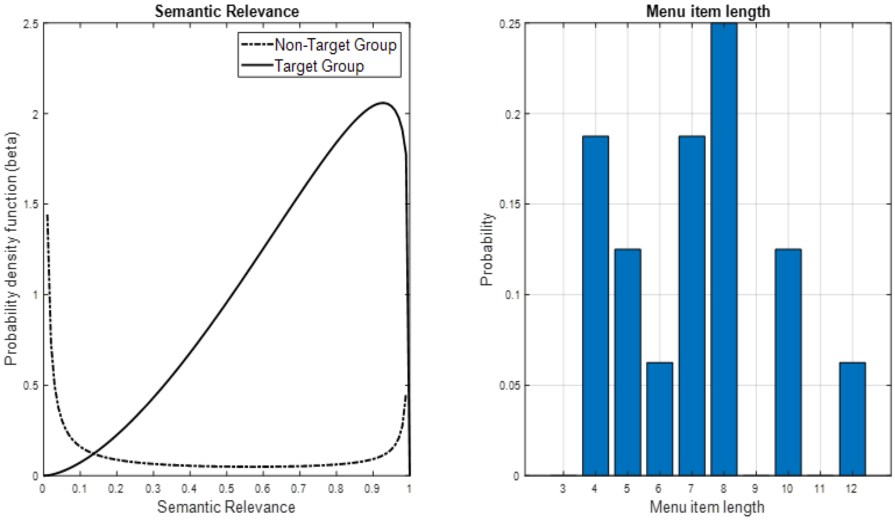


Figure 6.1: The ecological distribution of the semantic similarity (left panel) and the ecological distribution of menu items length (right panel).

After creating the ecological distributions of the menu search environment, the model was ready to be trained. The model was trained 20 million trials. The model was trained until performance converged. The performance convergence was realized when the average returns (rewards) of each of the last epochs were similar, see Figure 6.2. The training process produced a training file (3.08 GB).

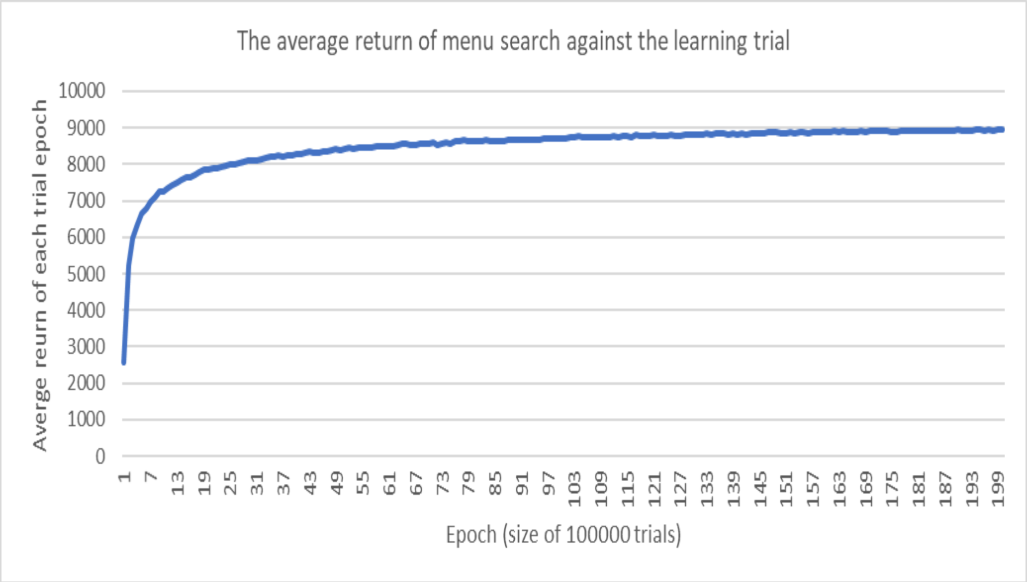


Figure 6.2: The average return of menu search against the learning trial.

Each trail involved training the model on a menu constructed by sampling randomly from the ecological distributions of semantic and shape relevance created before. The target location in this menu was randomly selected from the distribution of the target location.

An example of how a menu sample is constructed is given. First, the semantic relevance score of the target item is set to 1. Then, the semantic relevance scores of the distractor items are sampled from the ecological distributions in Figure 6.1, left panel. The distractor items can be within the target group or outside the target group. The semantic relevance scores of the distractor items within the target group are sampled from the 'Target Group' distribution, see Figure 6.1, left panel. The semantic relevance scores of the distractor in other groups are sampled from the 'Non-Target Group' distribution, see Figure 6.1, left panel.

In this study, 8-items menu samples were used instead of 16-items menu samples. This was to simplify the training process as no need to train the model on a high-performance computer as with 16-items menu samples.

After learning the menu search strategies through the training process, the model was ready to be tested to check whether it can predict outlying menu search performance due to menu semantics. To test the model's prediction, the model was given two menu search tasks. These tasks involve finding and selecting a specific target.

The first task was to search for a specific target in 100 newly generated 8-items menu samples, that all have the target in the same location (item 3). The target location (item 3) was chosen as a representative of the target location. The menu semantics of these generated menu samples were different as the semantic relevance scores of their items were sampled randomly from the ecological distributions in Figure 6.1, left panel. Some of them were semantically organised (the target group contains semantically similar items, and the non-target group contains semantically unrelated items), and some of them were poorly organised (the target group contains semantically unrelated items, or the non-target group contains semantically similar items).

The second task was the same as the first task except that the location of the target in the generated menu samples was (item 7).

In each menu search task, the model predicted the time spent to find the target for each menu sample. The model's prediction resulted from only exploiting the optimal policy.

6.2.2 Results

- **The model's prediction in the first menu search task**

The selection accuracy of the model was 100%. The model predicted the time spent to find the target for each menu sample (individual trial). The boxplot was used to display the distribution of the predicted menu search time of 100 trials and to identify outliers. Three outlying performance cases were found, see Figure 6.3. To find out whether these cases were due to poorly organized menus, the menu semantics of the menu samples that generated these cases were checked. Most menu samples (2 out of 3) were semantically organised, see Appendix D.1. This indicates that the predicted outlying menu search performance was not due to poorly organized menus.

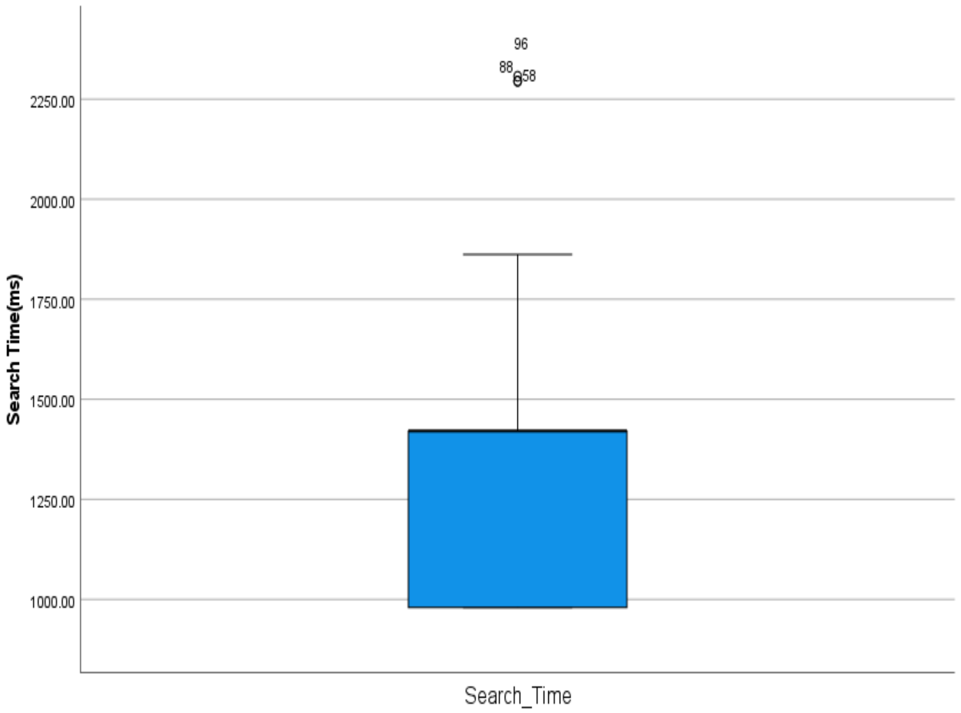


Figure 6.3: A Boxplot of the model's prediction of menu search time in the first menu search task.

- **The model's prediction in the second menu search task**

The selection accuracy of the model was 99%. The model predicted the time spent to find the target for each menu sample (individual trial). The boxplot was used to display the distribution of the predicted menu search time of 100 trials and to identify outliers. One outlying performance case was found, see Figure 6.4. To find out whether this case was due to poorly organized menus, the menu semantics of the menu sample that produced this case was checked. The menu sample was semantically organised, see Appendix D.2. This again indicates that the predicted outlying menu search performance was not due to poorly organized menus.

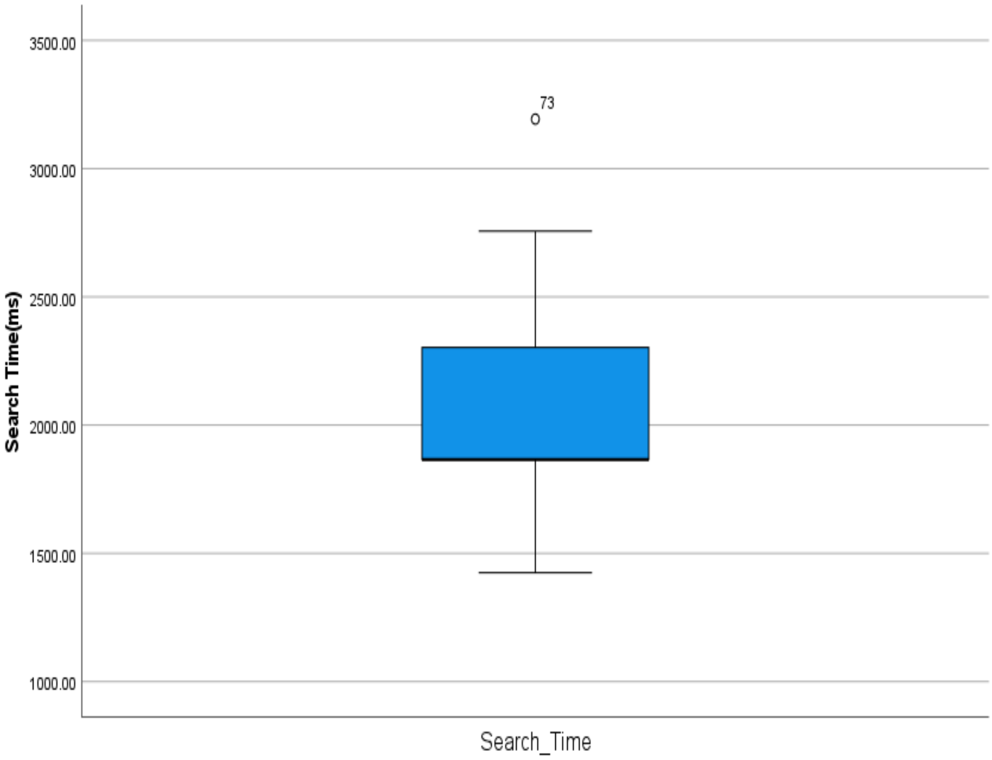


Figure 6.4: A Boxplot of the model's prediction of menu search time in the second menu search task.

- **Other observations**

When the same trained model was run two times using the same menu samples, each run generated different predictions compared with another run. For example, different outlying performance cases were predicted using the same trained model and the same menu samples, see Figure 6.5.

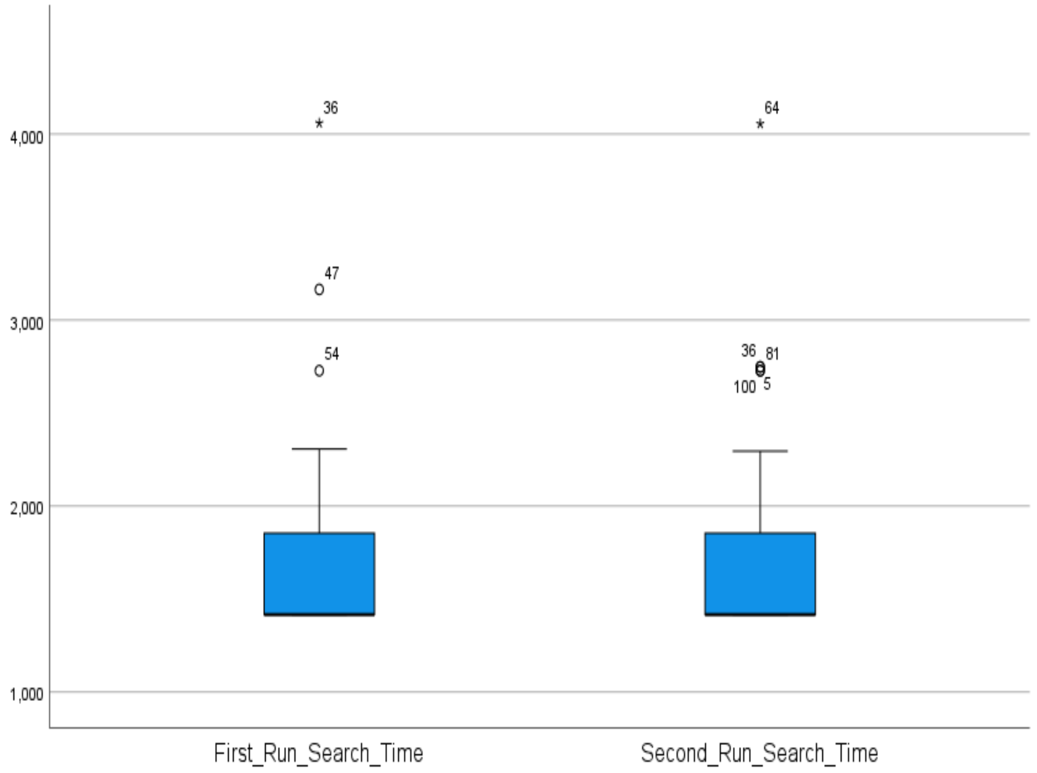


Figure 6.5: The search time prediction of the same trained model in two different runs using the same menu samples.

6.3 Training and Testing the Model Using Data Collected by the Card Sorting

Card sorting was used as another method to collect the semantic similarity ratings because it was interesting to use both methods to collect the semantic similarity ratings of menu items and see whether they introduce different features on data. This is important when populating a model based on this data.

In Study 4 (Chapter 5), the semantic similarity ratings were collected using the card sorting method. The collected semantic similarity ratings were prepared to be used in the training and testing model. The training and testing procedures are mentioned in the following section.

6.3.1 Method

Procedure

First, the menu search environment was created using the menu used in Study 2 (Chapter 4). This menu was used to determine the ecological distribution of menu item length, see Figure 6.5, right panel. Also, it was used to determine the ecological distribution of the semantic similarity ratings of menu items, see Figure 6.6, left panel. These semantic similarity ratings were collected from human participants in Study 4 (Chapter 5) using the card sorting method.

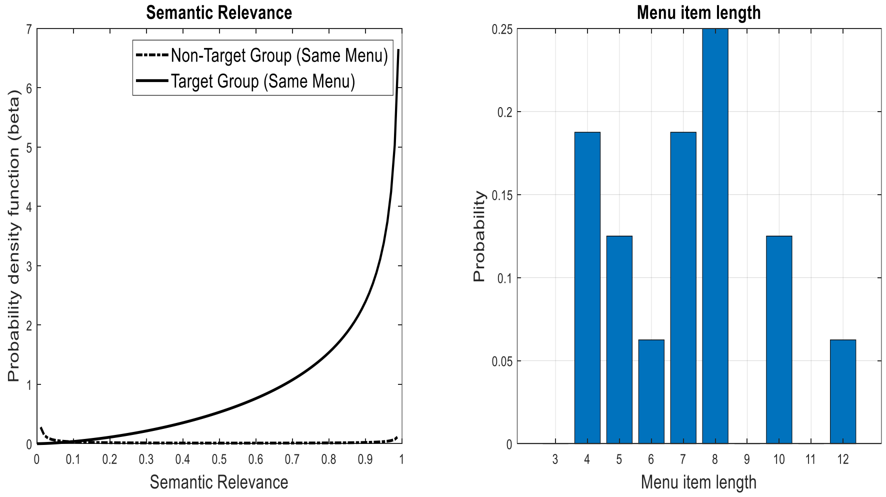


Figure 6.6: The ecological distribution of the semantic similarity ratings collected by the card sorting (left panel) and the ecological distribution of menu items length (right panel).

After that, the same training and testing procedures mentioned in the previous section 6.2.1 were applied exactly. The model was trained 20 million trials. The model was trained until performance converged. The performance convergence was realised when the average returns (rewards) of each of the last epochs were similar, see Figure 6.7. The training process produced a training file (2.03 GB).

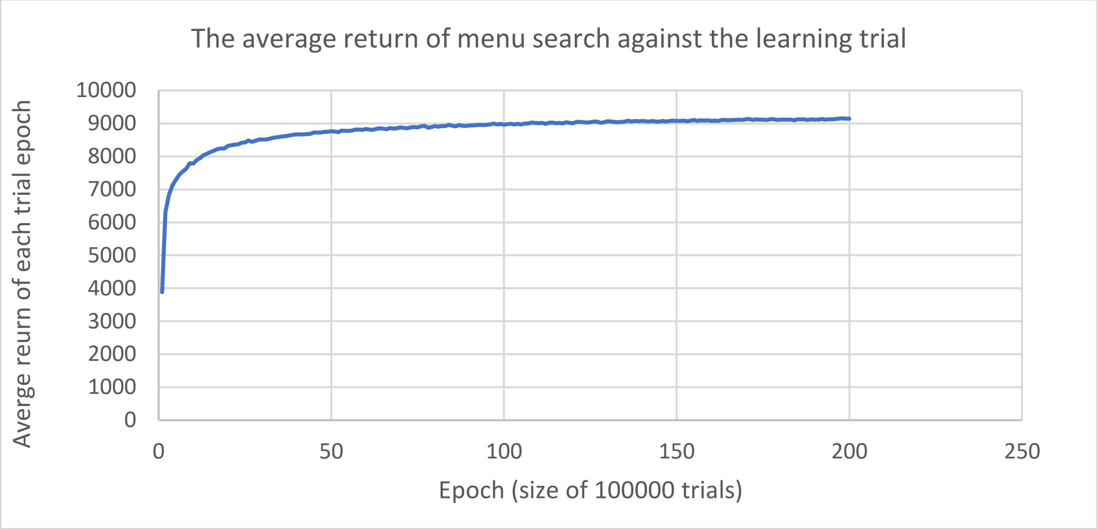


Figure 6.7: The average return of menu search against the learning trial (the model trained on card sorting data).

Following that, the model was tested by giving it two menu search tasks. The first task was to search for a specific target in 100 newly generated 8-items menu samples, that all have the target in the same location (item 3). However, the menu semantics of these menu samples were different. Some of them were semantically organised (the target group contains semantically similar items, and the not-target group contains semantically unrelated items), and some of them were poorly organised (the target group contains semantically unrelated items, or the not-target group contains semantically similar items).

The second task was the same as the first task except that the location of the target in the generated menu samples was (item 7).

In each menu search task, the model predicted the time spent to find the target for each menu sample. The model's prediction resulted from only exploiting the optimal policy.

6.3.2 Results

- **The model's prediction in the first menu search task**

The selection accuracy of the model was 96%. The model predicted the time spent to find the target for each menu sample (individual trial). The boxplot was used to display the distribution of the predicted menu search time of 100 trials and to identify outliers. Four outlying menu search performance cases were found, see Figure 6.8. To find out whether these cases were due to poorly organized menus, the menu semantics of the menu samples that produced these cases were checked. All menu samples were semantically organized, see Appendix D.3. This again indicates that the predicted outlying menu search performance was not due to poorly organized menus.

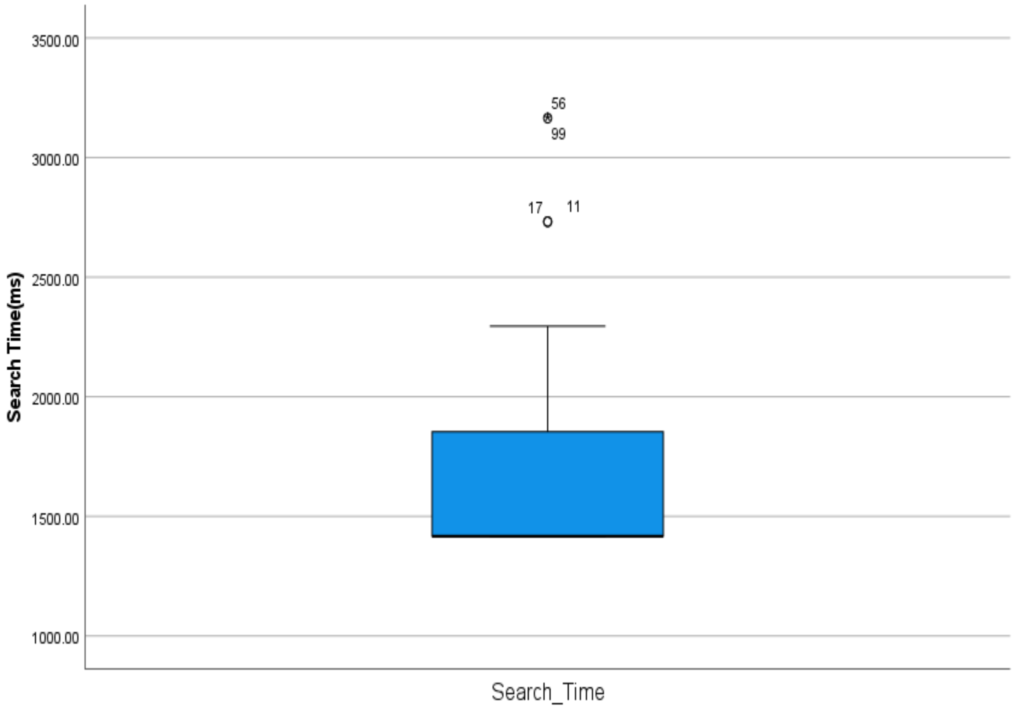


Figure 6.8: A Boxplot of the model's prediction of menu search time in the first menu search task (the model trained on card sorting data).

- **The model's prediction in the second menu search task**

The selection accuracy of the model was 97%. The model predicted the time spent to find the target for each menu sample (individual trial). The boxplot was used to display the distribution of the predicted menu search time of 100 trials and to identify outliers. Three outlying menu search performance cases were found, see Figure 6.9. To find out whether these cases were due to poorly organized menus, the menu semantics of the menu samples that produced these cases were checked. All the menu samples were semantically organized, see Appendix D.4. This again indicates that the predicted outlying menu search performance was not due to poorly organized menus.

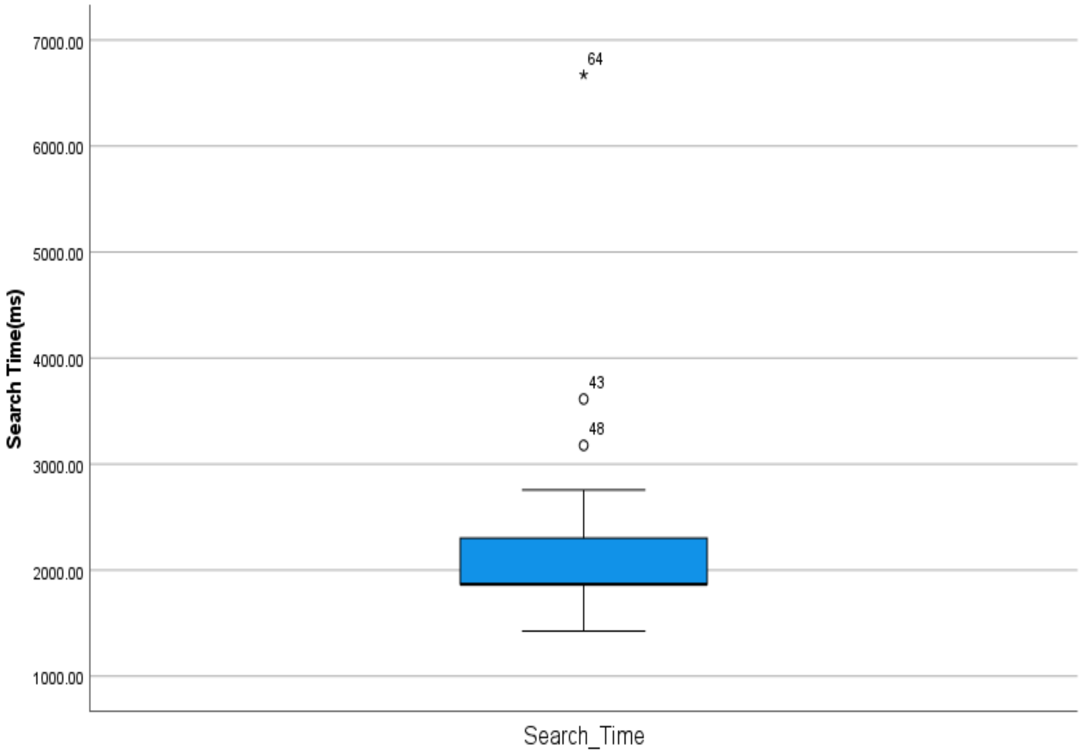


Figure 6.9: A Boxplot of the model's prediction of the menu search time in the second menu search task (the model trained on card sorting data).

6.4 Discussion

This study aimed to test whether the adopted menu search model can predict outlying menu search performance due to menu semantics. The model was trained and tested two times using different menu sets that were construed from different ecological distributions of semantic similarity ratings. The model was first trained and tested using data collected by the pairwise similarity ratings method. The trained model predicted outliers, but they were not due to menu semantics. After that, the model was trained and tested using data collected by the card sorting method. Again, the trained model predicted outliers, but they were not due to poorly organized menus. Therefore, this study found no evidence that menu semantics plays a role in outlying menu search performance.

The adopted menu search model aims to model menu search in general. It does not aim at modelling how people learn particular menus and the position of particular items. Therefore, it might be that it cannot be used to model an individual's menu search performance as it was done in this study.

Additionally, by comparing the adopted model with the TreeWalker model that was proposed by Schiller and Cairns (2008) and was able to predict outliers due to menu semantics, the TreeWalker model has a clean interface that uses semantic relevance only to determine the actions, while the adopted model in this study assumes that semantic and shape relevance determine the actions. Moreover, the adopted model assumes that there is uncertainty in visual information processing, therefore, the exact encoded values are prone to noise, as indicated by Chen et al. (2015). This might justify it was difficult in the adopted model to isolate the effect of menu semantics on individual menu search performance.

Although the Treealker model was able to predict outliers due to menu semantics, it could be that model was not explored enough once it produced a good result. The model was not fitted to experimental data. Therefore, more investigations are needed to check the role of menu semantics in outlying menu search performance.

Lastly, the adopted menu search model was not scalable. Training the model using 16-items menu samples needs substantial computation and produces a very large training file. Therefore, 8-items menu samples were used in this study. However, this is not an issue for the finding of this study because if the model is not able to predict outliers in searching 8-items menu samples, the model will not be able to predict outliers in searching menus of any length.

Chapter 7

Study 6: Investigating the Role of Individual Differences in Outlying Menu Search Performance

7.1 Introduction

Study 2 (Chapter 4) found several outlying menu search performance cases. Study 3 and Study 4 (Chapter 5) found that there are individual differences in perceiving the semantic similarity of menu items. However, as these two studies were conducted separately, it was not possible to relate the outlying menu search performance to people who perceive the semantic similarity of menu items differently. Yin (2018) built her work on this research and investigated whether participants who perceive the semantic similarity of menu items differently perform poorly in menu search tasks. Also, she investigated whether outlying menu search performance is due to specific individuals. She found that neither the semantic similarity perception of menu items nor the person was responsible for outlying performance.

Yin's study recruited 40 participants and collected the semantic similarity ratings using the pairwise similarity ratings method. It would be interesting to conduct a similar study with larger sample size and use the card sorting method to collect the semantic similarity ratings of menu items.

This study aimed to investigate the role of individual differences in outlying menu search performance. It checked whether having a different perception of menu semantics plays a role in outlying menu search performance. Additionally, it checked whether outlying menu search performance is due to specific participants.

7.2 Method

7.2.1 Design

The goal of this study was to check whether having a different perception of menu semantics plays a role in outlying menu search performance. Additionally, to check whether outlying menu search performance is due to specific participants.

This study consisted of two tasks: a card sorting task and a menu search task. In the card sorting task, participants' perception of menu semantics was elicited by using card sorting. In the menu search task, there were six menu search trials. In each trial, the participants were asked to search for a specific item in a two-level hierarchical menu, the first level presented the main menu categories, and the second level presented the menu items under each category. A two-level hierarchical menu design was used in this study because it was interesting to investigate outlying menu search performance using more complex menus that have more than one level. These types of menus would allow following completely incorrect paths, and this might increase the likelihood of outlying menu search performance.

Participants' menu search performance data (search time and path) was recorded in each trial. The search time is the time taken by participants to find and select the target item, and it was measured in seconds. The path shows where participants went during the search trial. Paths can be filtered into direct and indirect success. Direct success means that participants found the target item directly without moving back to the main categories. Indirect success means that participants found the target items but moved back to the main categories at least once.

Multiple menu search trials were designed to check whether the outlying menu search performance is caused by specific participants. If outlying performance is due to a permanent feature in a participant, the same participant will show outlying performance in all or most menu search trials.

The menu search trials order was not counterbalanced. They were presented to the participants in the same order. This allows checking the role of having a different perception of menu semantics. If participants do not carry out the trials in the same order, their knowledge of the menu might override their initial perception of menu semantics and consequently affect their menu search performance.

Presenting the menu search trials in the same order might introduce order effects such as practice effects, fatigue effects and boredom effects. Fatigue and boredom effects are not likely to happen as the task is short and less intense to do. Practice effects are more likely to happen as participants might become better as they become more familiar with the menu. However, this study did not aim to compare the participants performance in different menu search trials. It aimed just to compare the participants performance within an individual menu search trial to check whether their perception of menu semantics affects their menu search performance in this trial.

7.2.2 Participants

In this study, Amazon Mechanical Turk (MTurk) was used to recruit the participants. MTurk is a crowdsourcing Internet marketplace where researchers can recruit participants for their studies. This online crowdsourcing service has been used and validated as a tool for different kinds of research. As for the experimental cognitive science research, it was found by Crump et al. (2013) that the data collected by the MTurk was fairly high quality and comparable to the quality of the data collected in the controlled lab.

Using MTurk in this study enabled having a wider and more diverse population. The wider and more diverse population might help in investigating the role of individual differences in outlying menu search performance.

To ensure the quality of the collected data, participation was restricted to workers who have greater than or equal to 5000 approved Human Intelligence Task (HITs) and who have a HITs approval rate of greater than 95%.

One hundred and one (101) participants were recruited. Fifty-nine (59) participants were male, and forty-two (42) participants were female. They belonged to different age groups, see Figure 7.1. Their educational level ranged between high school and postgraduate degrees. Most participants have a good experience in purchasing electronic devices online, but the rest of the participants have little experience, see Figure 7.1.

Each participant was paid USD 4.00 (equivalent to GBP 3.26) for participation in this study. This payment was calculated based on a recommendation by Burleigh (2019) that workers should be paid an amount equal to a minimum hourly wage. He stated that if the task takes an average of one hour to complete, the payment can be USD 7.25-12 to every worker completing the task. Since the estimated time to complete the tasks in this study was 10 minutes or less, four dollars seemed a fair payment.

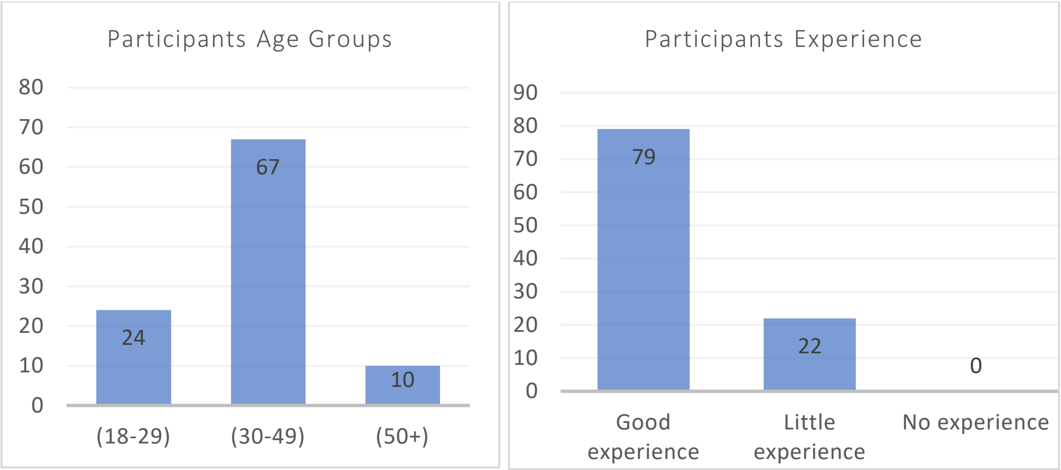
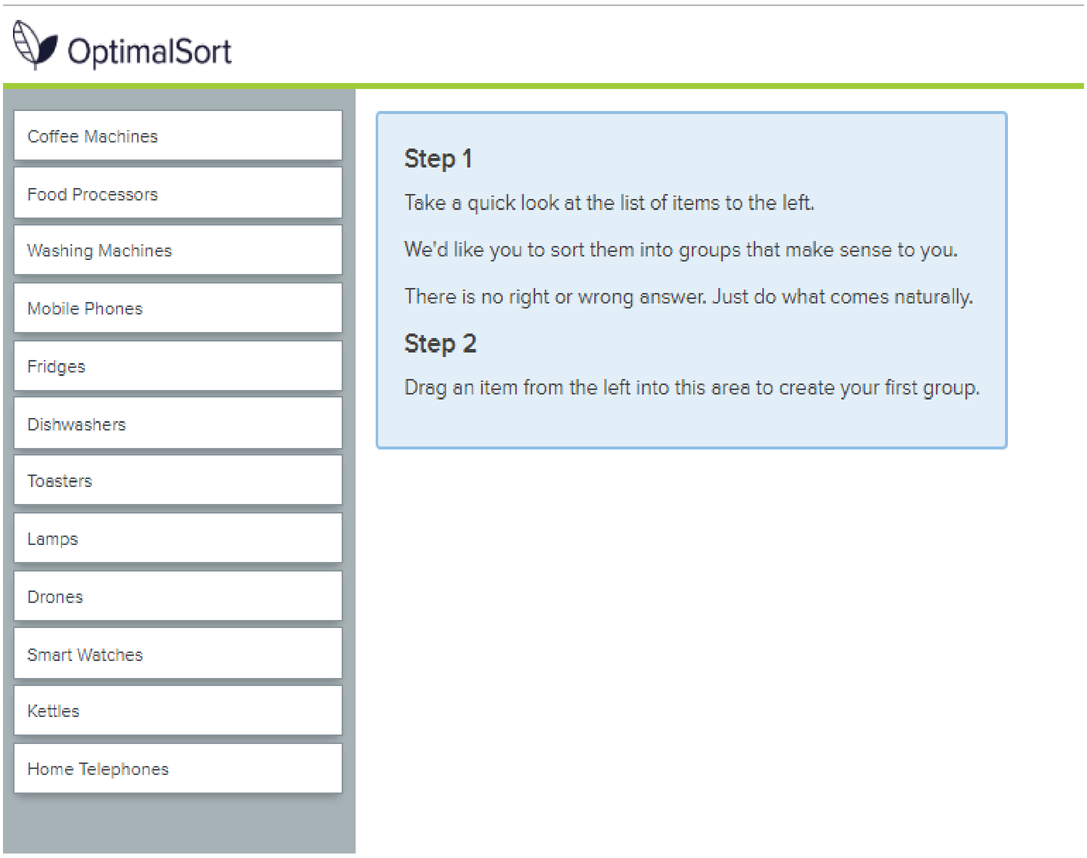


Figure 7.1: Demographics of participants.

7.2.3 Tasks and Materials

Participants were asked to perform two tasks: the card sorting task and the menu search task. In the card sorting task, participants were asked to sort the presented items into groups of similar items. In the menu search task, participants were asked to carry out six menu search trials. In each trial, participants were asked to click on a specific target item on a hierarchical menu (two-levels menu) as quickly and as accurately as they could. The menu search trial starts once the button “Start task” is clicked and the menu appears on the screen. The menu search trial finishes when participants select the target item.

The menu items that were used in this study were the same menu items used in Yin's (2018) study. These menu items were derived from the John Lewis online store's menu, more specifically, from the electricals top-level category. The electricals category was chosen because nowadays most people rely on electricals in their daily life. And as new electronic devices are regularly introduced, the menu items under the electricals category differ across different e-commerce websites (Yin, 2018). The electricals category was simplified in this study to include three subcategories; Home Appliance, Small Appliance and Smart Tech, while there are six subcategories in the original category on John Lewis online store. Four items were chosen under each subcategory (12 items in total). Under Home appliance, there were Washing machines, Fridges, Dishwashers and Lamp. Under Small Appliances, there were Coffee machines, Kettles, Toasters and Food processors. Under Smart Tech, there were Mobile phones, Smart watches, Drones and Home telephones. These twelve selected menu items were used to create the cards that should be sorted by the participants in the card sorting task, see Figure 7.2.



Step 1

Take a quick look at the list of items to the left.

We'd like you to sort them into groups that make sense to you.

There is no right or wrong answer. Just do what comes naturally.

Step 2

Drag an item from the left into this area to create your first group.

Figure 7.2: The card sorting task (created by OptimalSort).

Also, the three selected subcategories, Home Appliance, Small Appliance and Smart Tech and their items were used to construct the two-level hierarchical menu used in the menu search task, see Figure 7.3. The target items selected for the six menu search trials were as follow: Kettles, Home Telephones, Lamps, Coffee Machines, Dishwashers and Drones (two items from each subcategory).

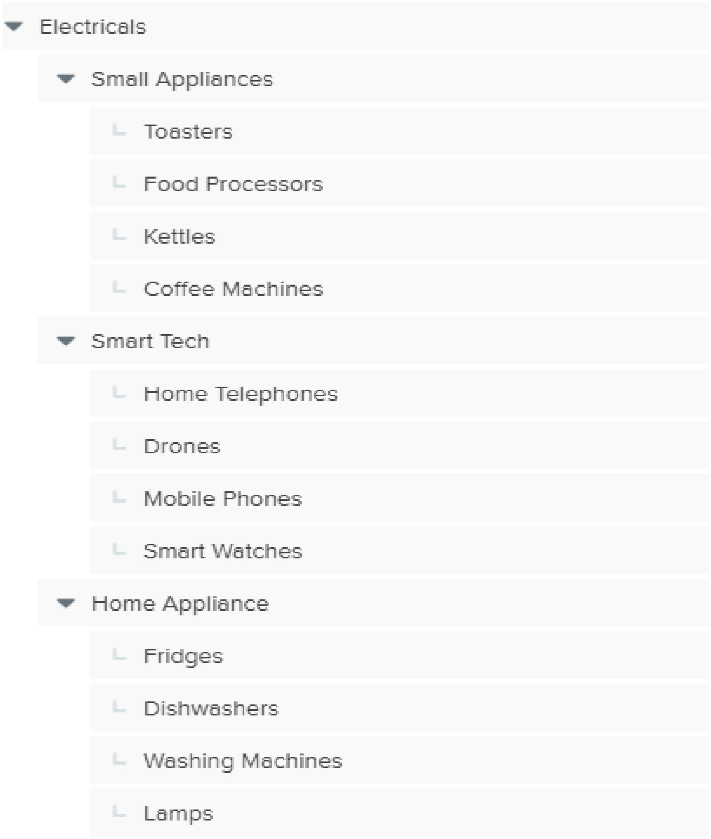


Figure 7.3: The menu used in the menu search task.

This study used Optimal Workshop- a user research platform that offers several tools such as card sorting and tree testing. The card sorting tool "OptimalSort" was used to create the card sorting task. The tree testing tool "Treejack" was used to create the menu search task. This tool automatically records participants' performance data, such as task completion time and path. Also, it allowed presenting the menu alone on the website page. This helped in avoiding any source of distractions that can be found on real websites and affect user performance.

The OptimalSort tool presented the twelve cards to the participants, see Figure 7.2. These cards represented menu items. To sort these items, participants need to drag and drop each item with similar items.

The Treejack tool presented a menu to the participants. This menu has three subcategories that were organised in a hierarchical structure, see Figure 7.3. This menu was folded in the formal task, and to see the items belong to a specific subcategory, participants need to click on this subcategory, see Figure 7.4.

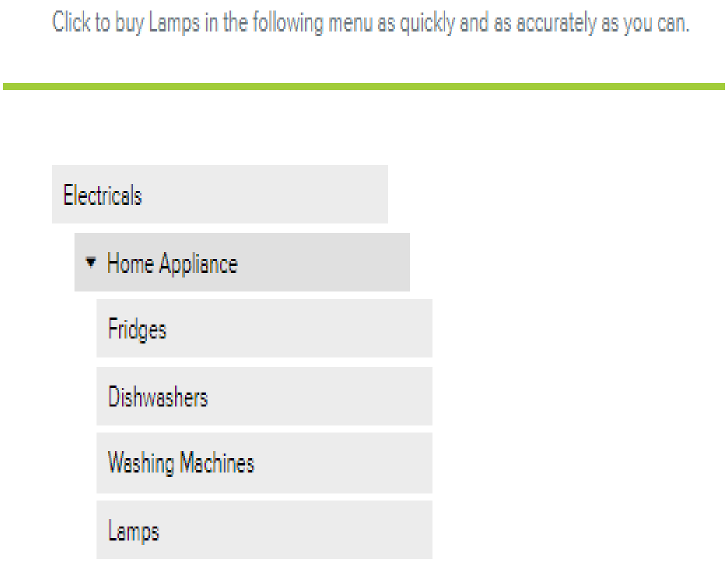

Click to buy Lamps in the following menu as quickly and as accurately as you can.

Figure 7.4: An example of a menu search trial (created by Treejack tool).

7.2.4 Procedure

This study was unmoderated and done online using an Amazon Mechanical Turk (MTurk) website. The description of this study was presented to the MTurk Workers once they clicked on the HIT link of this study on the HITs page on the MTurk website. If they accepted participating in this study, clicking on the link would redirect them to this online study created by the Optimal workshop platform. Once they clicked on the link, the study information sheet was presented to explain the whole information about this study, see Appendix E.1. Then, they were asked to sign the consent form by ticking the box to indicate their agreement to take part in this study, see Appendix E.1. After this, they were asked to fill out a short survey about their demographics and their experience in purchasing electricals online. Then, they were redirected to the card sorting task link. Once they finished this task, they were redirected to the menu search task link. This task consists of 6 trials as well as one practice trial at the beginning. After that, they went back to the MTurk website to confirm the successful completion of this study. The average time spent to complete the whole study was 3.5 minutes.

7.2.5 Pilot

This study was piloted to ensure that everything was clear and understandable. The links of this study were sent to two colleagues, and they were asked to complete this study and give feedback. Based on their feedback, there was only one slight change recommended. One colleague has suggested modifying one of the demographic questions that were about the experience in purchasing electricals online. This closed question offered two responses (yes/no). It was suggested that the yes and no responses be replaced with three responses that represented three possible levels of experience in purchasing electricals online. Therefore, three responses were set to this question. These responses were as follows: I have a good experience, I have a little experience, and I have no experience.

7.2.6 Data Analysis

The responses of MTurk workers should be checked to detect the signs of subject inattentiveness as recommended by Cheung et al. (2017). After checking the responses, eight participants seemed inattentive. They did not follow the instructions as they sorted the cards randomly in the card sorting task or selected the wrong menu item in the menu search task. The responses of these participants were removed from the collected data. After removing these responses, the total number of responses was ninety-three (93).

The responses to the card sorting task were recorded by the OptimalSort tool. This tool automatically generated a similarity matrix. Each similarity value in the similarity matrix represents the percentage of participants who place these two cards in the same group. This similarity matrix was downloaded as an Excel sheet. Then, it was then analysed using the same clustering techniques (hierarchical clustering and MDS) that were used in Study 3 and Study 4 (Chapter 5). These clustering methods are used to find out the semantic groups of menu items according to participants sorting that represents their perception of menu semantics.

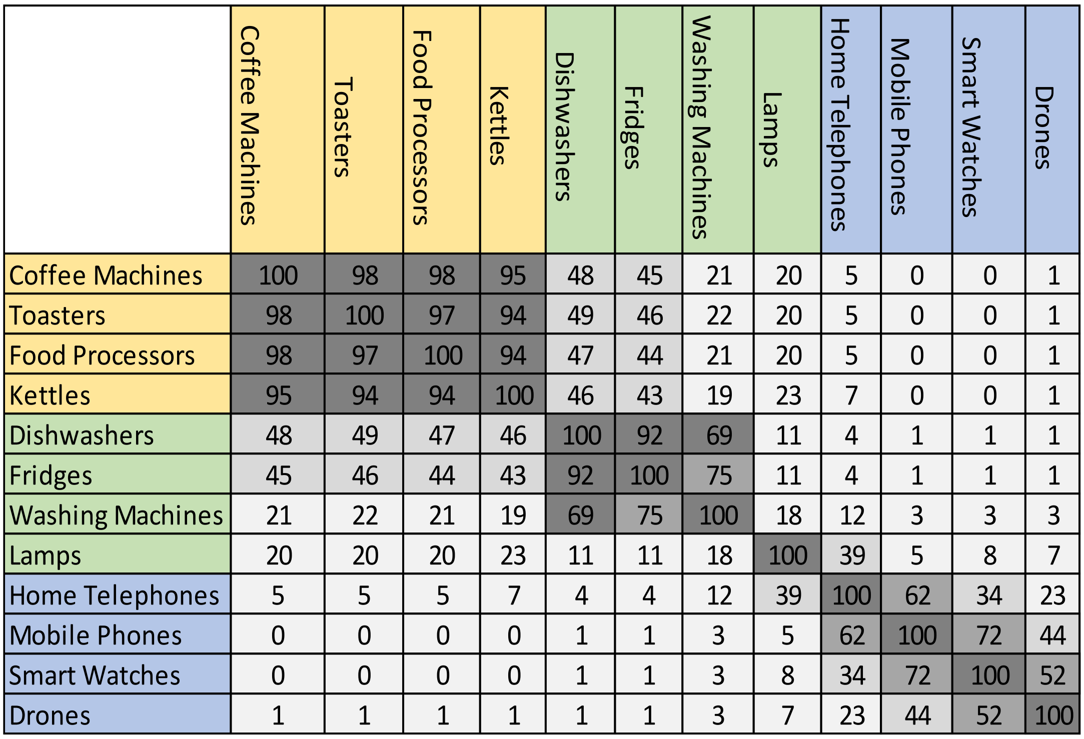
Additionally, to find out the similarity between the original menu sort and participants' sorts, minimum edit distance was used. The minimum edit distance measures how far two card sorts are by calculating the minimum number of changes to transform one sort into another sort (Deibel and Anderson, 2005). One change involves moving one card from one group to another one.

Participants’ performance in the menu search trials was recorded by the Treejack tool. This tool provided detailed results of each menu search trial. These results included the time taken by each participant to complete the menu search trial (search time) and the path followed by each participant to find and select the target item. These results were downloaded as Excel files for analysis. The boxplots were used to display the distribution of search time for each menu search trial and to identify outliers in each menu search trial. After identifying the outlying menu search performance cases, these cases were studied in detail to find out whether they are due to having a different perception of menu semantics or due to specific participants.

7.3 Results

7.3.1 The perception of menu semantics

By looking at the generated similarity matrix, see Figure 7.5, it was clear that Coffee machines, Toasters, Food processors and Kettles were perceived as highly similar to each other as they were grouped by a high percentage of participants. Also, Dishwashers, Fridges and Washing machines were grouped by most participants. Lamps and Home telephones were grouped by 39 % of participants. Home telephones and Mobile phones were also believed to be similar to each other. Mobile phones and Smart watches were grouped by 72% of participants. Smart watches and Drones were thought similar to each other by about half of the participants.



	Coffee Machines	Toasters	Food Processors	Kettles	Dishwashers	Fridges	Washing Machines	Lamps	Home Telephones	Mobile Phones	Smart Watches	Drones
Coffee Machines	100	98	98	95	48	45	21	20	5	0	0	1
Toasters	98	100	97	94	49	46	22	20	5	0	0	1
Food Processors	98	97	100	94	47	44	21	20	5	0	0	1
Kettles	95	94	94	100	46	43	19	23	7	0	0	1
Dishwashers	48	49	47	46	100	92	69	11	4	1	1	1
Fridges	45	46	44	43	92	100	75	11	4	1	1	1
Washing Machines	21	22	21	19	69	75	100	18	12	3	3	3
Lamps	20	20	20	23	11	11	18	100	39	5	8	7
Home Telephones	5	5	5	7	4	4	12	39	100	62	34	23
Mobile Phones	0	0	0	0	1	1	3	5	62	100	72	44
Smart Watches	0	0	0	0	1	1	3	8	34	72	100	52
Drones	1	1	1	1	1	1	3	7	23	44	52	100

Figure 7.5: A 12-by-12 similarity matrix, the similarity values represented the percentage of participants who place these two items in the same group.

The hierarchical clustering technique was applied to the similarity matrix. Four hierarchical clustering methods (single, complete, average and ward.D) were used. Four clusters (semantic groups) were found by using the average and single hierarchical clustering methods, see figures 7.6 (a) and (b):

Cluster 1: Lamps;

Cluster 2: Home Telephones, Drones, Mobile Phones and Smart Watches;

Cluster 3: Kettles, Food Processors, Coffee Machines and Toasters;

Cluster 4: Washing Machines, Dishwashers and Fridges.

The complete and ward.D hierarchical clustering methods generated four clusters (semantic groups), see figures 7.6 (c) and (d):

Cluster 1: Drones, Mobile Phones and Smart Watches;

Cluster 2: Lamps and Home Telephones;

Cluster 3: Kettles, Food Processors, Coffee Machines and Toasters;

Cluster 4: Washing Machines, Dishwashers and Fridges.

The multidimensional scaling (MDS) method found four clusters (semantic groups); see figure 7.7.

Cluster 1: Lamps;

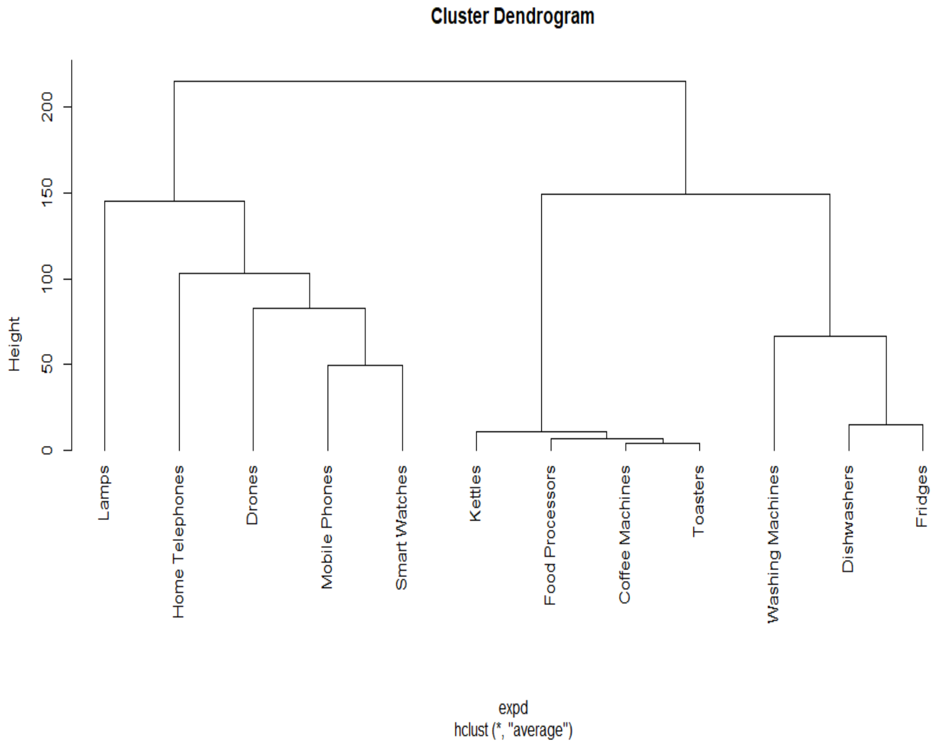
Cluster 2: Home Telephones, Drones, Mobile Phones and Smart Watches;

Cluster 3: Kettles, Food Processors, Coffee Machines and Toasters;

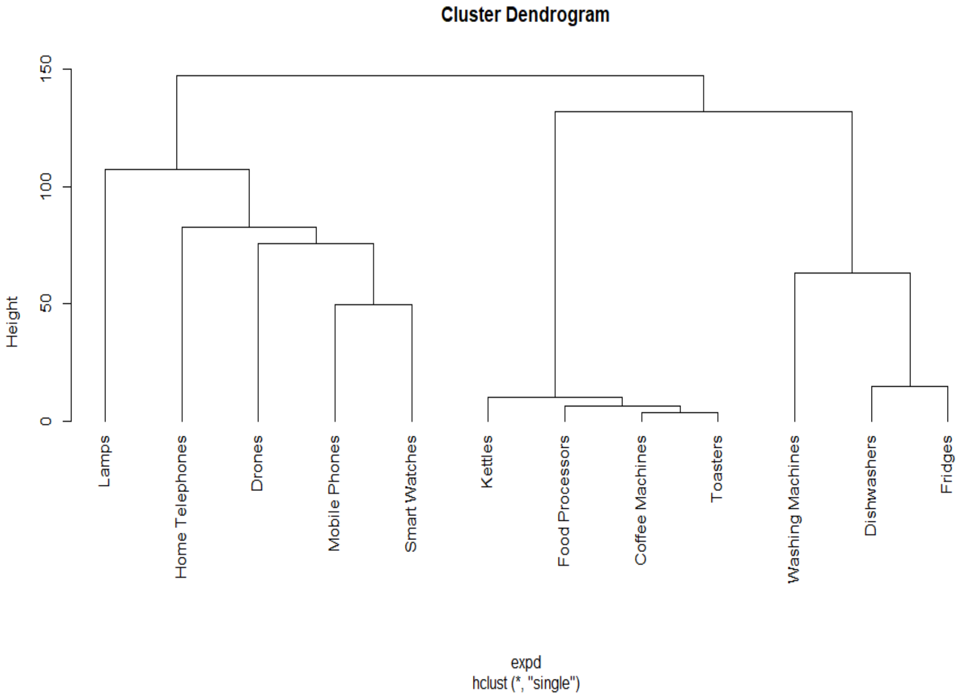
Cluster 4: Washing Machines, Dishwashers and Fridges

Again, Lamps was isolated in one cluster as found by the average and single hierarchical clustering methods.

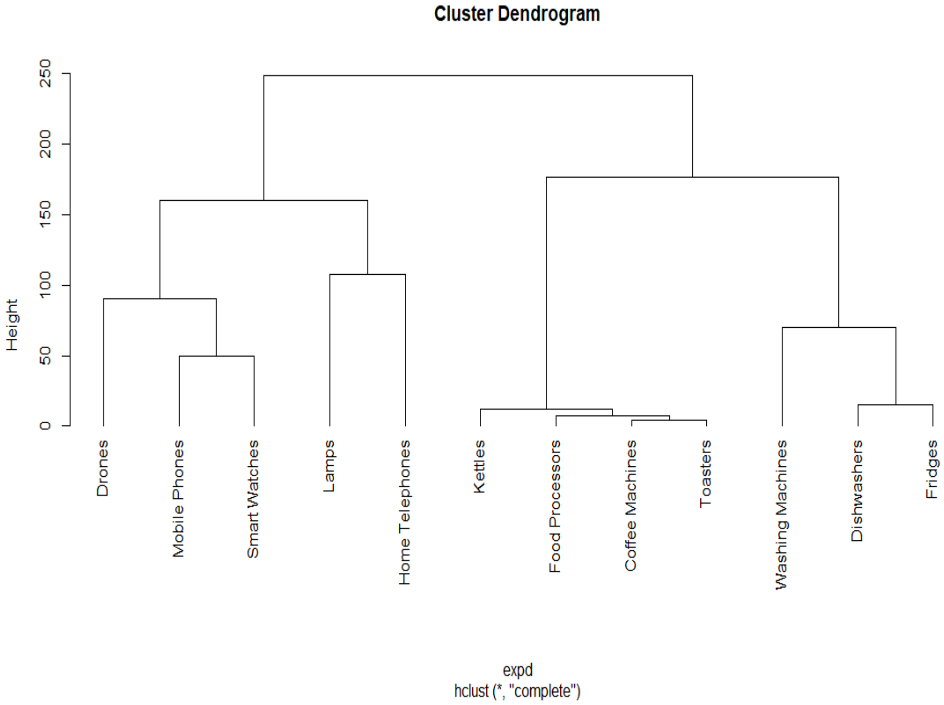
By comparing the results of the four hierarchical clustering methods (average, single, complete, ward.D) and the MDS method, three robust clusters were produced by all methods. These clusters fit the original menu categories (Smart Tech, Small Appliances and Home Appliances). However, the Lamps item was found in a separate cluster, see Figure 7.6. This might indicate that the Lamps item confused most participants as they did not group it with the similar items in its original group. Therefore, it appeared isolated from its similar items in all clustering diagrams. Home Telephones item might also confuse some participants as they did not group it with the similar items in its original group. Therefore, it appeared in a different cluster as in complete and ward.D hierarchical clustering diagrams.



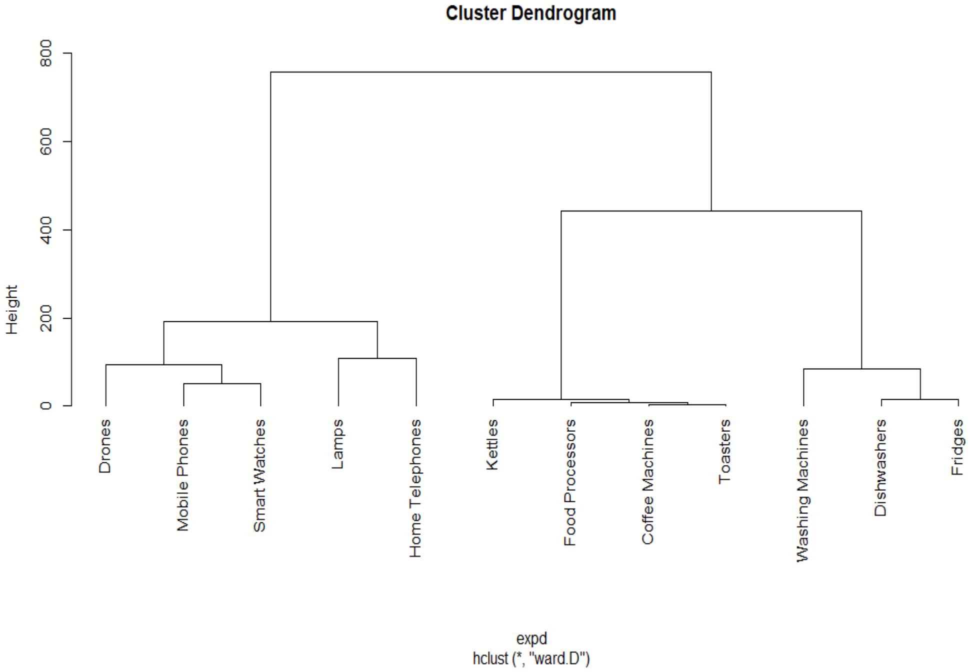
(a) average



(b) single



(c) complete



(d) ward.D

Figure 7.6: The clusters of menu items generated by the four hierarchical clustering methods (average, single, complete, ward.D).

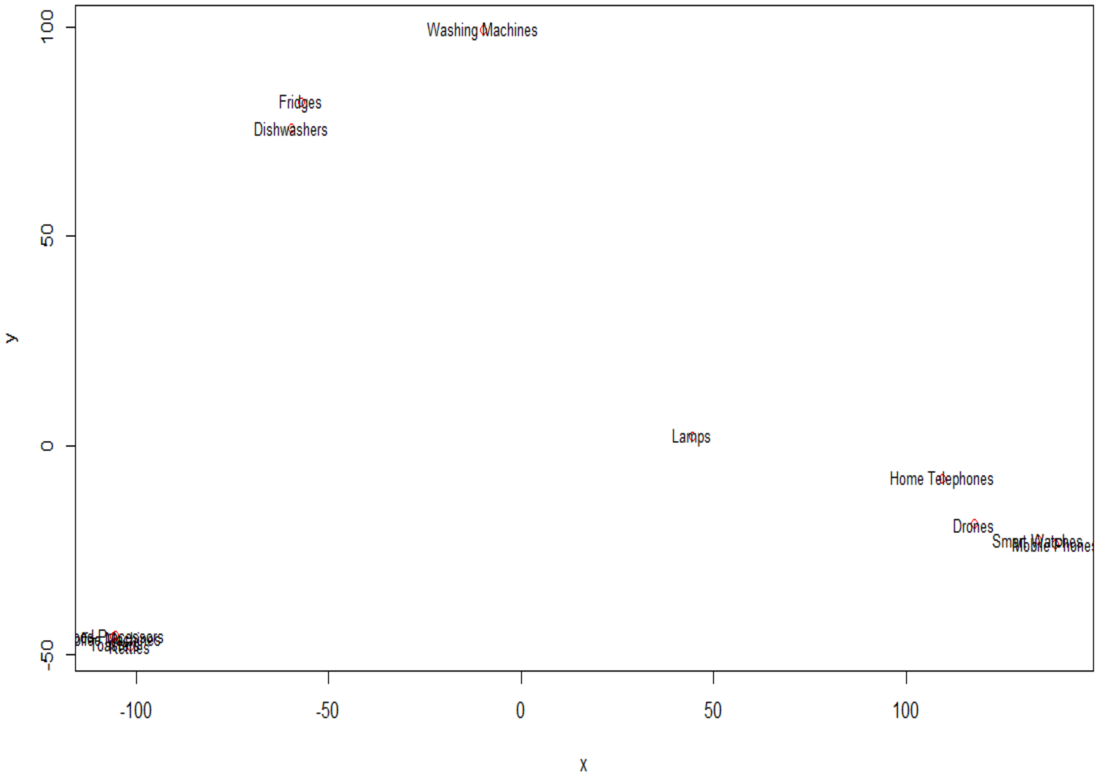
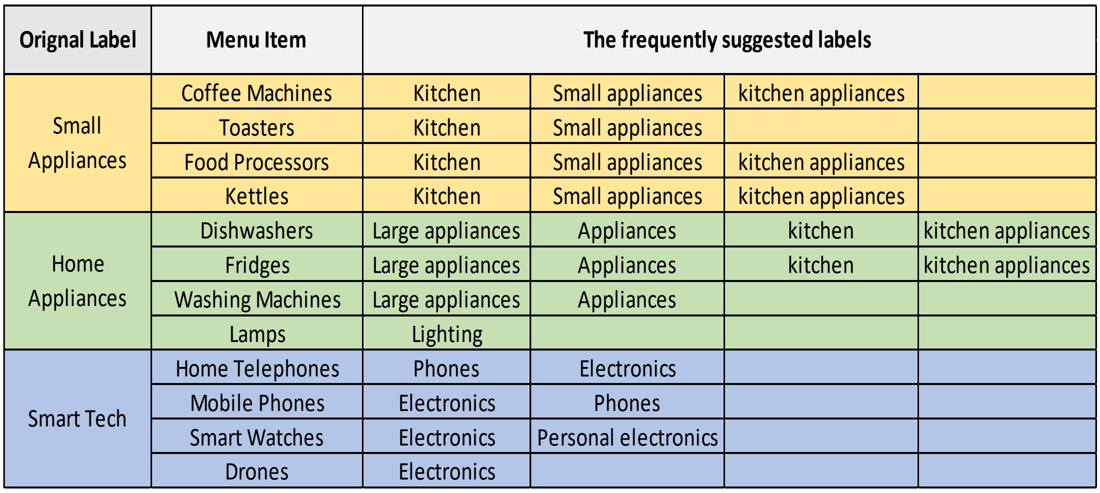


Figure 7.7: The clusters of menu items generated by the multidimensional scaling (MDS).

With regard to the suggested labels for the created groups, it was found that different labels were suggested by the participants to name the created groups. However, some labels were frequently suggested, see Table 7.1. The frequently suggested labels mean that were suggested by 10 participants or more. Kitchen and Small appliances labels were used frequently to name the menu items group that belonged to Small Appliances subcategory. The menu items that belonged to Home appliances subcategory were categorised under Large appliances and Appliances labels, except Lamps item that was repeatedly placed under the Lighting label. Electronics and Phones labels were used frequently to name the menu items group that belonged to Smart Tech subcategory.

The frequently suggested labels were not similar to the original subcategories' labels found in John Lewis online store's menu, except for Small appliances that matched the original label.



Orignal Label	Menu Item	The frequently suggested labels			
Small Appliances	Coffee Machines	Kitchen	Small appliances	kitchen appliances	
	Toasters	Kitchen	Small appliances		
	Food Processors	Kitchen	Small appliances	kitchen appliances	
	Kettles	Kitchen	Small appliances	kitchen appliances	
Home Appliances	Dishwashers	Large appliances	Appliances	kitchen	kitchen appliances
	Fridges	Large appliances	Appliances	kitchen	kitchen appliances
	Washing Machines	Large appliances	Appliances		
	Lamps	Lighting			
Smart Tech	Home Telephones	Phones	Electronics		
	Mobile Phones	Electronics	Phones		
	Smart Watches	Electronics	Personal electronics		
	Drones	Electronics			

Table 7.1: The frequently suggested labels for each item.

7.3.2 The average menu search performance

The average search time and the percentage of the direct success (finding the target item without backtracking) for each menu search trial were presented in Table 7.2.

Trial#	Average search time	Direct success percentage
Trial#1(select Kettles)	(μ= 5.9, σ= 3.78)	(65.6 %)
Trial #2 (select Home Telephones)	(μ= 8.74, σ= 3.75)	(33.3%)
Trial#3 (select Lamps)	(μ= 6.48, σ= 3.38)	(50.5%)
Trail#4 (select Coffee Machines)	(μ= 5.44, σ= 2.43)	(63.44%)
Trial#5 (select Dishwashers)	(μ= 4.14, σ= 1.7)	(84.9%)
Trial#6 (select Drones)	(μ= 3.49, σ= 1.6)	(93.5%)

Table 7.2: The average search time and the percentage of direct success for each menu search trial.

Trial 2 has the longest average search time and lowest percentage of direct success, followed by trial 3. Trial 6 has the shortest average search time and the highest percentage of direct success, followed by trial 5. This might be because of the learning effect, the participants learned the menu, and therefore they found the target items directly.

7.3.3 The outlying menu search performance

Outliers were found in each trial, see Figure 7.8. The total number of outlying menu search performance cases was 26 cases produced by 20 participants, 3 cases in trial 1, one case in trial 2, 2 cases in trial 3, 4 cases in trial 4, 7 cases in trial 5 and 9 cases in trial 6.

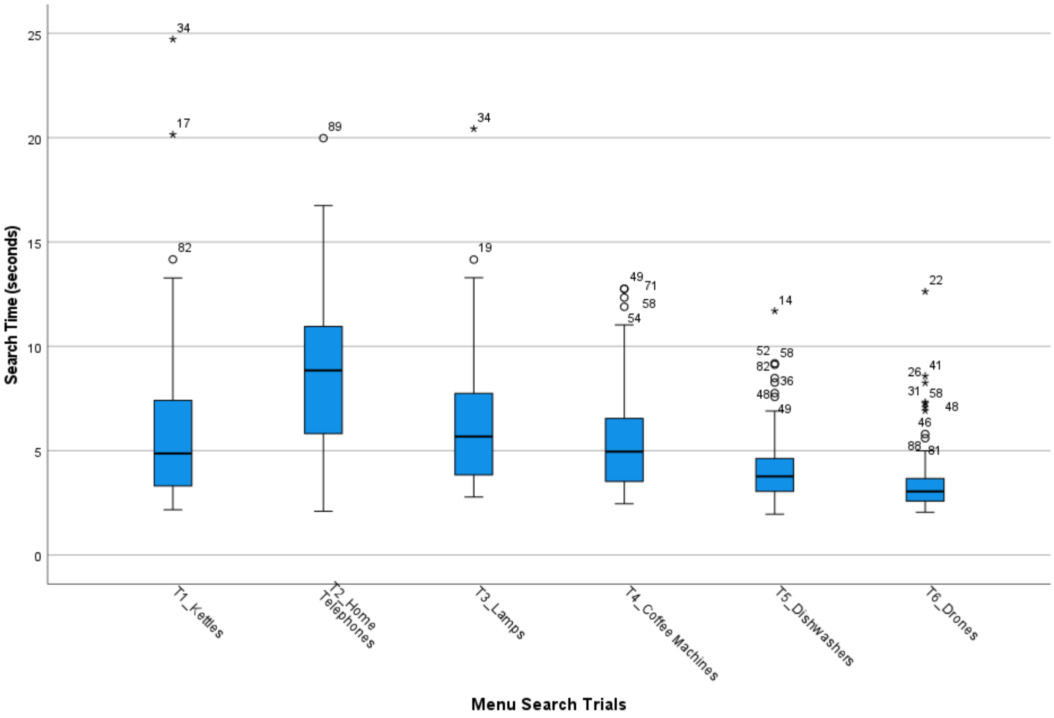


Figure 7.8: Boxplots of search time for each menu search trial.

The outlying menu search performance cases were analysed to find out whether these cases were due to having a different perception of menu semantics. Also, to check whether these cases were due to specific participants. Additionally, other factors such as age and experience were considered.

- **The effect of having a different perception of menu semantics**

The minimum edit distance was calculated to find out the similarity of each participant’s sort to the original menu sort, see Figure 7.9. The average distance was (μ= 3.42, σ= 1.56), see Figure 7.10.

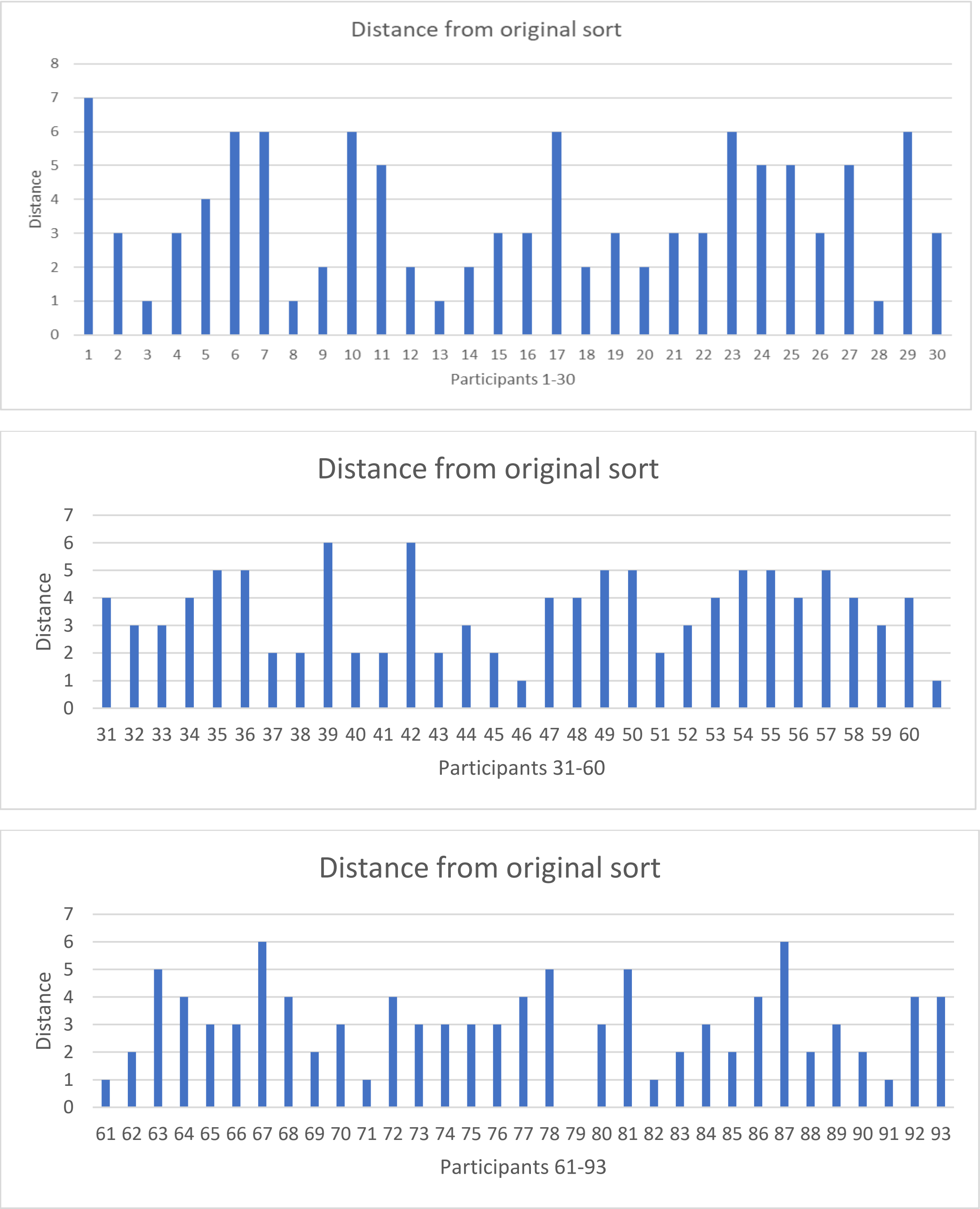


Figure 7.9: The distance of participants from the original menu sort.

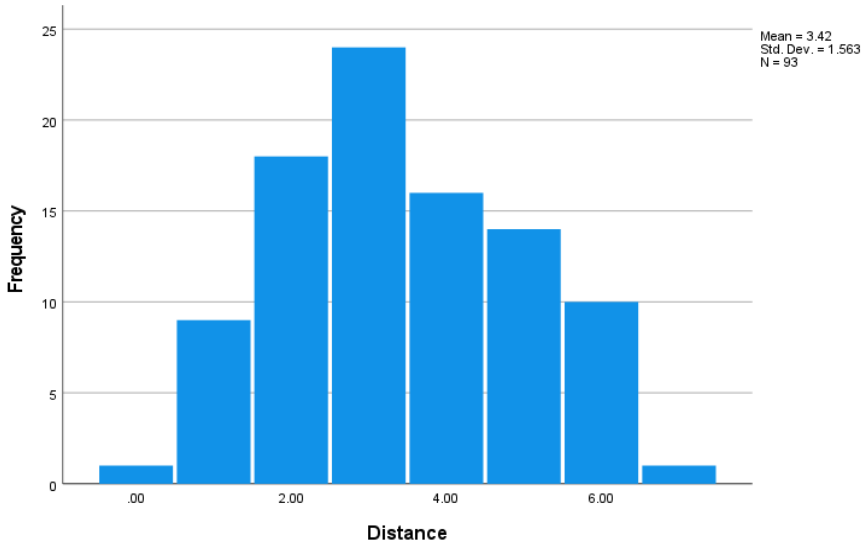


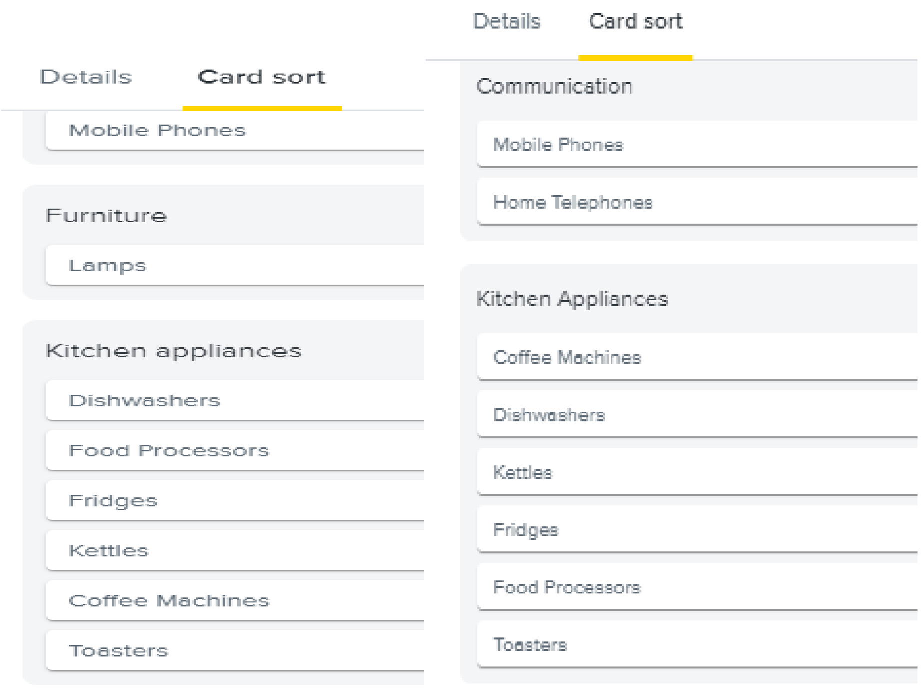
Figure 7.10: The histogram of the distance of participants.

To find out whether outliers sorted the menu items differently, the distance of each outlier was checked, see Table 7.3. Their distances were not different compared with the distances of non-outliers, see Figure 7.9 and Figure 7.10.

Additionally, the card sorting of each outlier in the first three menu search trials was checked. It was found that outliers and non-outliers sorted the menu items in a similar way. Examples of outliers' card sorting are given. Outlier (P 17) occurred in the first trial (select Kettles). This outlier grouped Kettles item with similar items, see Figure 7.11 (a). This was just like many participants who are not outliers and put these items together, see Figure 7.11 (b). Outlier (P 89) occurred in the second trial (select Home Telephones). This outlier could not group Home telephones item with similar items, see Figure 7.12, just like other participants did not put Home telephones item with the similar items as found in section 7.3.1. Also, outlier (P 19) occurred in the third trial (select Lamps). This outlier could not group Lamps item with the similar items, see Figure 7.13, just like other participants did not put Lamps item with the similar items as found in section 7.3.1.

The card sorting of the outliers in the last three menu search trials was not checked. This is because the speed up in the performance would suggest that the menu was learnt and, therefore, the participant's knowledge of the menu overrode their initial semantic perception.

Based on the abovementioned results, no relationship was found between having a different perception of menu semantics and outlying menu search performance.



Mobile Phones

Furniture
- Lamps

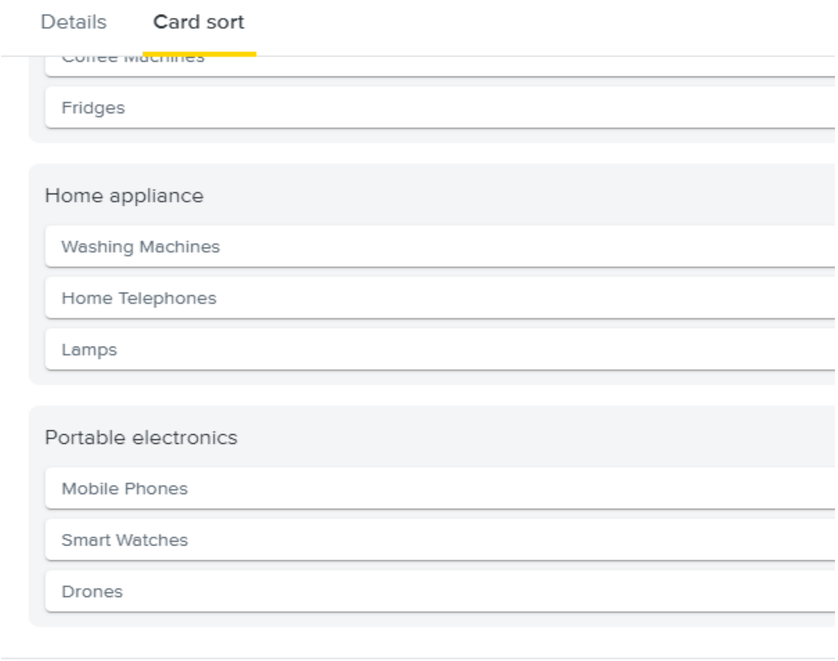
Kitchen appliances
- Dishwashers
- Food Processors
- Fridges
- Kettles
- Coffee Machines
- Toasters

Communication
- Mobile Phones
- Home Telephones

Kitchen Appliances
- Coffee Machines
- Dishwashers
- Kettles
- Fridges
- Food Processors
- Toasters

(a) (b)

Figure 7.11: (a) Outlier (P 17)'s card sorting, (b) Not outlier participant's card sorting.



Coffee Machines

Fridges

Home appliance

Washing Machines

Home Telephones

Lamps

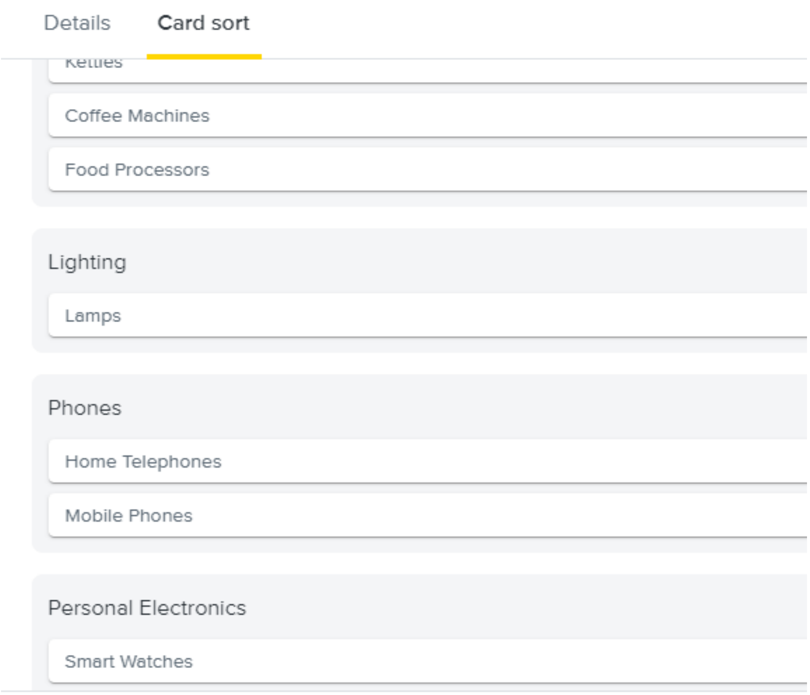
Portable electronics

Mobile Phones

Smart Watches

Drones

Figure 7.12: The outlier (P 89)'s card sorting. This outlier occurred in the second trial (select Home telephones).



Kettles

Coffee Machines

Food Processors

Lighting

Lamps

Phones

Home Telephones

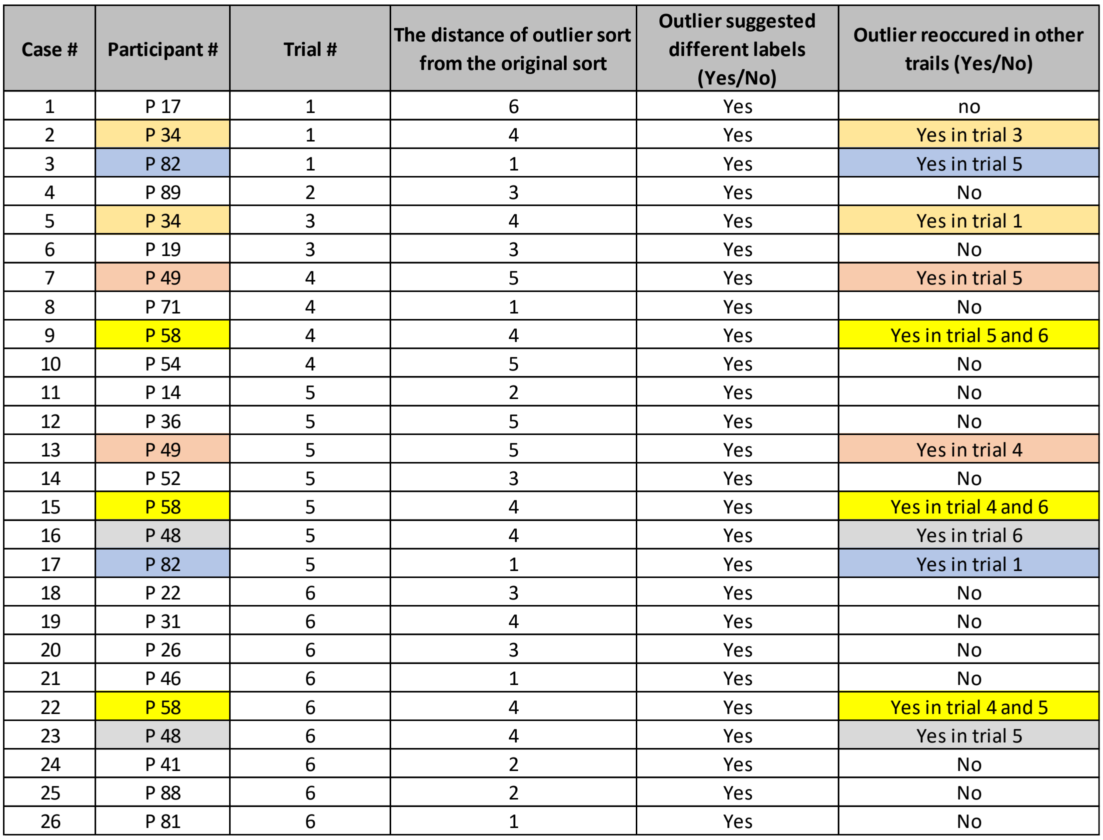
Mobile Phones

Personal Electronics

Smart Watches

Figure 7.13: The outlier (P 19)'s card sorting. This outlier occurred in the third trial (select Lamps).

As for the suggested labels, it was found that all outliers suggested different labels to name the created groups, see Table 7.3. These suggested labels were different compared with the original labels. For example, Kitchen was used by most outliers to label the kitchen related items. However, most average participants in this study also suggested different labels for the created groups. Therefore, there was no relationship between having different expectations of category labels and outlying menu search performance.



Case #	Participant #	Trial #	The distance of outlier sort from the original sort	Outlier suggested different labels (Yes/No)	Outlier reoccured in other trails (Yes/No)
1	P 17	1	6	Yes	no
2	P 34	1	4	Yes	Yes in trial 3
3	P 82	1	1	Yes	Yes in trial 5
4	P 89	2	3	Yes	No
5	P 34	3	4	Yes	Yes in trial 1
6	P 19	3	3	Yes	No
7	P 49	4	5	Yes	Yes in trial 5
8	P 71	4	1	Yes	No
9	P 58	4	4	Yes	Yes in trial 5 and 6
10	P 54	4	5	Yes	No
11	P 14	5	2	Yes	No
12	P 36	5	5	Yes	No
13	P 49	5	5	Yes	Yes in trial 4
14	P 52	5	3	Yes	No
15	P 58	5	4	Yes	Yes in trial 4 and 6
16	P 48	5	4	Yes	Yes in trial 6
17	P 82	5	1	Yes	Yes in trial 1
18	P 22	6	3	Yes	No
19	P 31	6	4	Yes	No
20	P 26	6	3	Yes	No
21	P 46	6	1	Yes	No
22	P 58	6	4	Yes	Yes in trial 4 and 5
23	P 48	6	4	Yes	Yes in trial 5
24	P 41	6	2	Yes	No
25	P 88	6	2	Yes	No
26	P 81	6	1	Yes	No

Table 7.3: The analysis of 26 outlying menu search performance cases.

- **The effect of a participant's trait**

The performance of outliers in each menu search trial was checked to find out some indications of their trait that might play a role in their poor performance, such as slowness and difficulty in recovering from mistakes.

To check whether outliers are slow people, the number of occurrences as outliers has been checked, see Table 7.3. Five participants have occurred more than once as outliers, four of them have occurred two times, and one has occurred three times. However, no one has occurred in all six trials. Also, the search time in the first trial and last trial were plotted and a Pearson correlation test was applied, see Figure 7.14.

There was a moderate correlation between the outliers' performance in the first trial and last trial, r = -0.446, r² = 0.199, see Figure 7.14. The negative correlation suggests that outliers are faster on the last trial if they take longer on the first trial. Therefore, there isn't a general effect of their trait but rather an indication that the initial processing of the menu leads to efficiencies later.

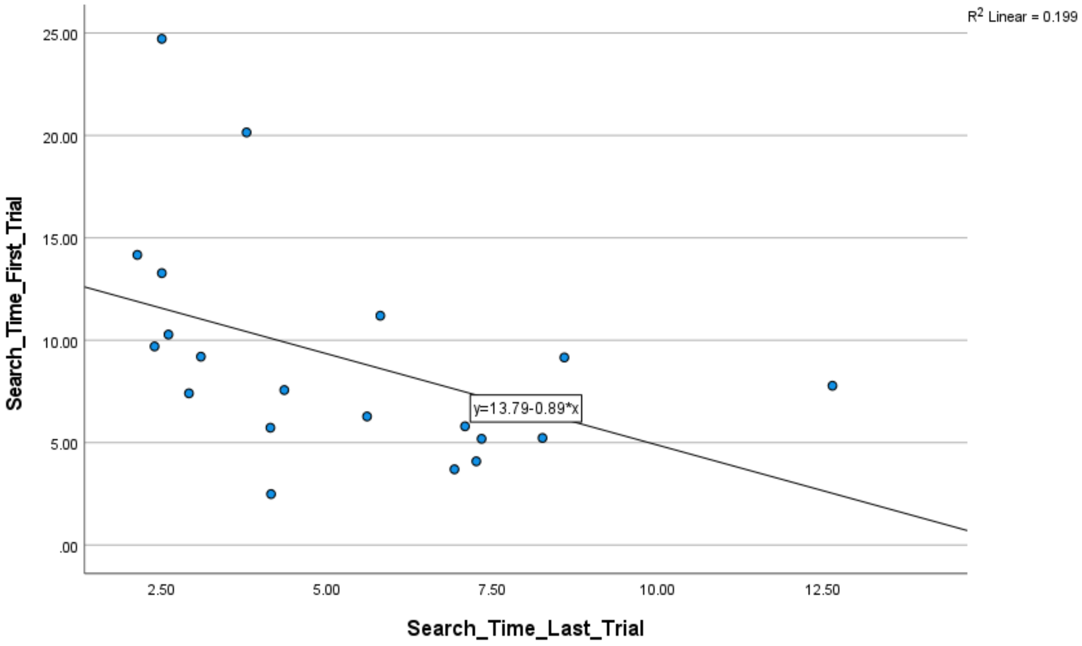
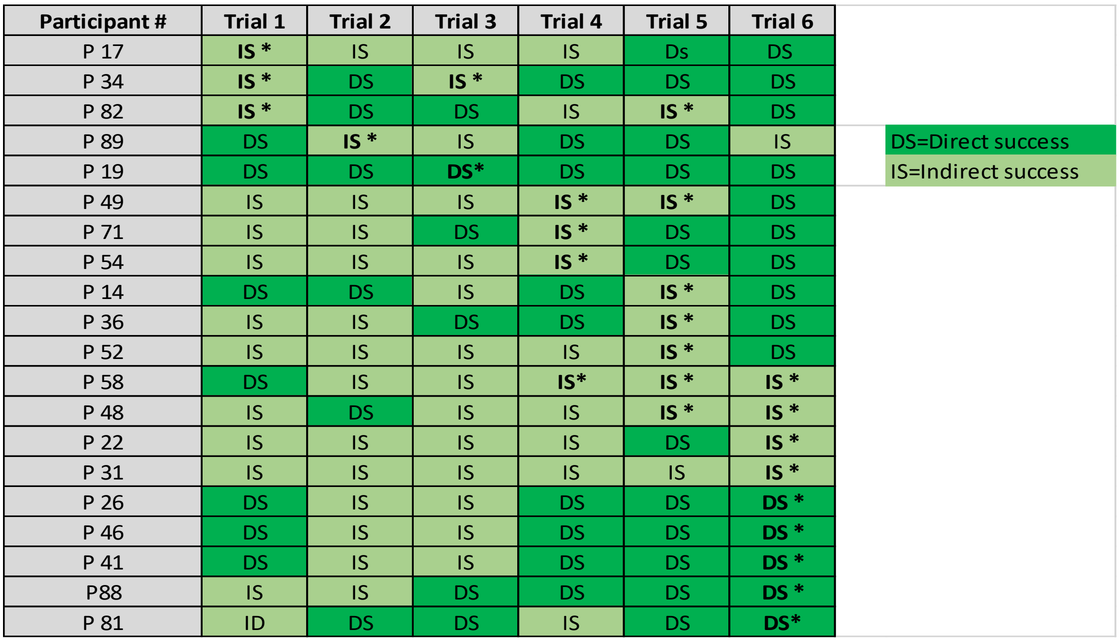


Figure 7.14: Search time for the first trial and the last trial.

To check whether outliers have difficulty in recovering from mistakes, their performance when they follow the wrong path and backtrack from it was checked, see Table 7.4. The followed path was classified into two types; direct success and indirect success. Direct success means that the participant went directly to the target item. Indirect success means that the participant did not go directly to the target item; they moved back through the menu subcategories before finding the target item. When checking the following path in each trial for each outlier, see Table 7.4, it was found that outliers have no problem in recovery after following the wrong path because, in some trials, the outliers followed the wrong path (indirect success) and did not show outlying performance. This indicates that outlying menu search performance was not related to the difficulty in recovering from choosing the wrong subcategory.

The abovementioned results indicate that no relationship was found between outlying menu search performance and a participant's trait.



Participant #	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
P 17	IS *	IS	IS	IS	Ds	DS
P 34	IS *	DS	IS *	DS	DS	DS
P 82	IS *	DS	DS	IS	IS *	DS
P 89	DS	IS *	IS	DS	DS	IS
P 19	DS	DS	DS*	DS	DS	DS
P 49	IS	IS	IS	IS *	IS *	DS
P 71	IS	IS	DS	IS *	DS	DS
P 54	IS	IS	IS	IS *	DS	DS
P 14	DS	DS	IS	DS	IS *	DS
P 36	IS	IS	DS	DS	IS *	DS
P 52	IS	IS	IS	IS	IS *	DS
P 58	DS	IS	IS	IS*	IS *	IS *
P 48	IS	DS	IS	IS	IS *	IS *
P 22	IS	IS	IS	IS	DS	IS *
P 31	IS	IS	IS	IS	IS	IS *
P 26	DS	IS	IS	DS	DS	DS *
P 46	DS	IS	IS	DS	DS	DS *
P 41	DS	IS	IS	DS	DS	DS *
P88	IS	IS	DS	DS	DS	DS *
P 81	ID	DS	DS	IS	DS	DS*

DS=Direct success

IS=Indirect success

Table 7.4: The followed path in each trial for each outlier (outlying performance occurred in this trial)*

- **The effect of age and experience**

Age and experience were also checked as factors that might play a role in the outlying performance cases in this study. About half the outliers (55%) belonged to the 30-49 age group. And most outliers (70%) have good experience in purchasing electricals online. This matches the whole participants' demographics percentages as most participants belonged to the 30-49 age group and have good experience in purchasing electricals online. The percentage of outlying performance in the older participants (50+) is 33% which is to some extent close to 27% of the percentage of outlying performance in younger participants (18-29), see Figure 7.15.

As for the influence of the experience in purchasing electricals online, the percentage of the outlying menu search performance in participants with good experience is 19%, and the percentage of the outlying menu search performance in participants with little experience is 31%, see Figure 7.16.

The percentages of outlying menu search performance in older participants and participants with little experience were a bit higher than the percentages of outlying menu search performance in younger participants and participants with good experience. This might indicate that age and experience are factors in outlying user performance.

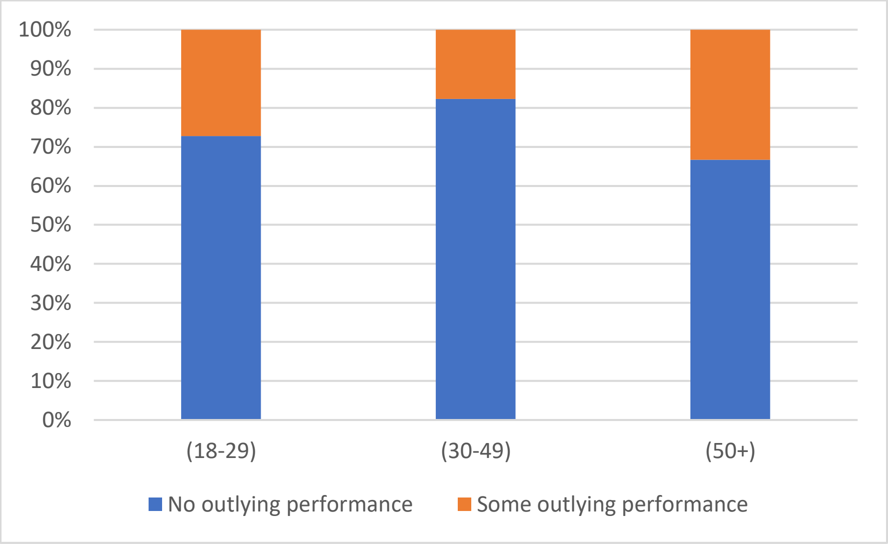


Figure 7. 15: The percentages of the participants with outlying performance against age groups.

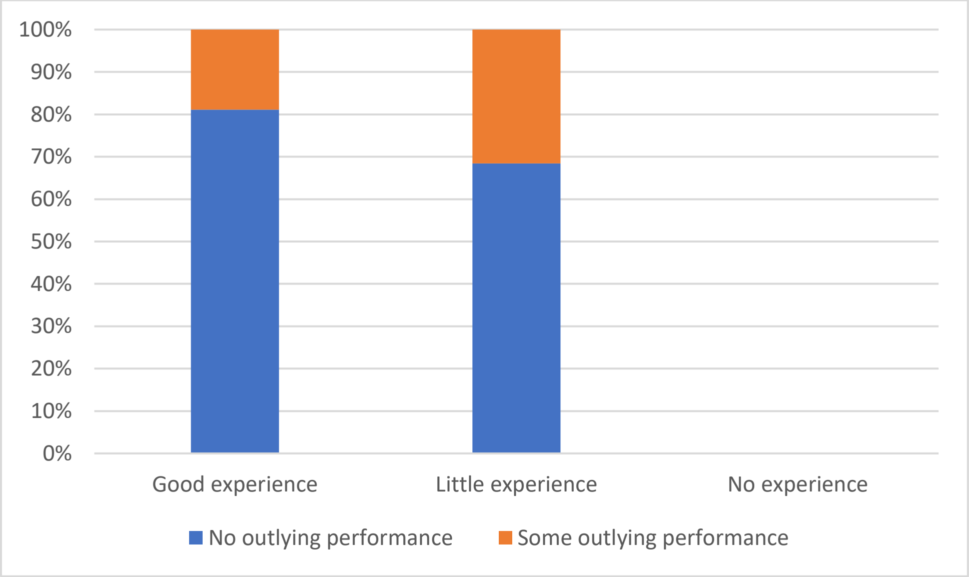


Figure 7.16: The percentages of the participants with outlying performance against experience.

To further investigate the influence of age and experience, the oldest participants (50+) and participants with little experience in purchasing electricals online were excluded from the sample. Twenty-four participants were excluded from the sample. The sample was analysed again. It was found that 24 outlying performance cases have occurred; 20 cases were the same cases that were found before, excluding the oldest participants and participants with little experience, see Figure 7.17. This indicates that other factors, not age and experience, play a role in outlying performance. The 24 outlying performance cases were analysed, see Table 7.5. The same results were obtained. The outlying menu search performance cases were not due to having a different perception of menu semantics. Also, these cases were not due to specific participants.

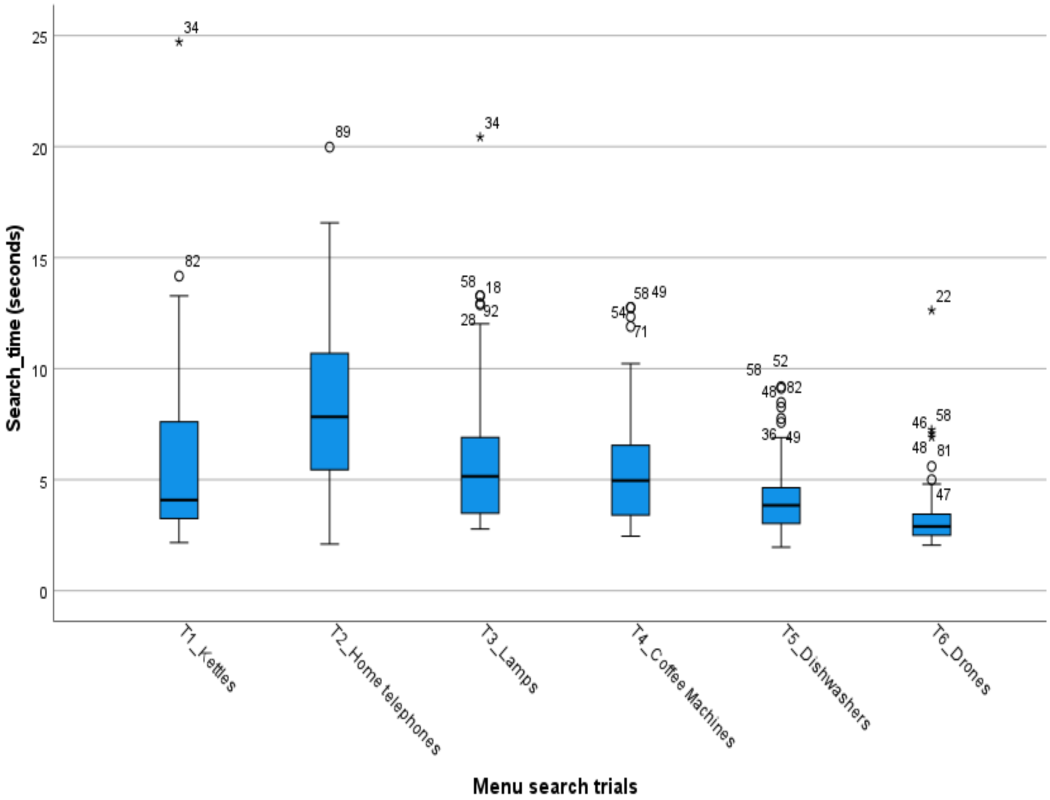


Figure 7.17: Boxplots of search time in each trial (after excluding the oldest participants and participants with little experience).

Case #	Participant #	Trial #	The distance of outlier sort from the original sort	Outlier suggested different labels (Yes/No)	Outlier reoccured in other trails (Yes/No)
1	P 34	1	4	Yes	Yes in trial 3
2	P 82	1	1	Yes	Yes in trial 5
3	P 89	2	3	Yes	No
4	P 34	3	4	Yes	Yes in trial 1
5	P 18	3	2	Yes	No
6	P 28	3	1	Yes	No
7	P 58	3	4	Yes	Yes in trial 4,5 and 6
8	P 92	3	4	Yes	No
9	P 49	4	5	Yes	Yes in trial 5
10	P71	4	4	Yes	No
11	P 58	4	4	Yes	Yes in trial 3, 5 and 6
12	P 54	4	5	Yes	No
13	P 36	5	5	Yes	No
14	P 49	5	5	Yes	Yes in trial 4
15	P 52	5	3	Yes	No
16	P 58	5	4	Yes	Yes in trial 3, 4 and 6
17	P 48	5	4	Yes	Yes in trial 6
18	P 82	5	1	Yes	Yes in trial 1
19	P 22	6	3	Yes	No
20	P 58	6	4	Yes	Yes in trial 3,4 and 5
21	P 48	6	4	Yes	Yes in trial 5
22	P 46	6	1	Yes	No
23	P 47	6	4	Yes	No
24	P 81	6	1	Yes	No

Table 7.5: The analysis of 24 outlying menu search performance cases (after excluding the oldest participants and participants with little experience).

- **Other observations**

The increased number of outliers in the last two menu search trials might be because of the increased skewness that comes with reduced menu search time with each trial. The reduction in average menu search time was because the participants got faster with each trial as they learnt the menu.

7.4 Discussion

This study aimed to investigate the role of individual differences in the outlying menu search performance. It checked whether having a different perception of menu semantics plays a role in outlying menu search performance. Also, it checked whether outlying menu search performance is due to specific participants.

The results of this study show no evidence that outlying menu search performance is related to having a different perception of menu semantics. The outliers have the same perception of menu semantics compared with other participants. Additionally, the results show no evidence that the occurred outlying menu search performance cases are due to specific individuals. The outliers did not show outlying performance on a regular basis. The outliers’ performance varied through the menu search trials. In some trials, they performed poorly, and in some, they did very well. These results confirm the findings of Yin (2018) that showed no evidence that outlying performance is due to having a different perception of menu semantics or due to specific individuals.

Furthermore, in trying to explain the occurred outlying menu search performance, age and experience were also investigated as possible factors that might play a role in outlying menu search performance. Outliers were found in all age groups. Also, outliers have different levels of experience in purchasing electricals online. However, the percentage of outlying performance in the older participants was a bit higher than in younger participants. Likewise, the percentage of outlying performance in participants with little experience was higher than the percentage of outlying performance in participants with good experience. By excluding the oldest participants and participants with little experience from the sample, most outlying performance cases still arise. This might indicate that age and experience are factors in the outlying user performance, but they are not the only factors, as was found by Auskerin (2012) in his work to investigate the really poor performance in usability tests.

The overall results of the participants' menu search performance support Schiller and Cairns’ (2008) observation that there is always one user or more who perform(s) poorly compared with other users. Outliers were found in each menu search trial.

With regards to the average menu search performance, it seemed to be influenced by the participants' perception of menu semantics. For example, most participants in the card sorting task were not able to group Home telephones and Lamps items with similar items as in the original groups. And when they were asked to search for these items in the menu search task, they took a long time to find and select them compared with other items. However, it could be that the long search time to find Home telephones and Lamps items was because of following the wrong path (visiting the wrong menu category) by most participants, and that affected the average menu search time. For example, when they were asked to search for Home telephones item, a lot of participants visited the Home Appliances category first because they thought that

Home telephones item is under this category; then, they backtracked and visited the Smart Tech category that contains Home telephones item. Also, when they were asked to search for Lamps item, about half participants visited the Small Appliances category first because they thought that Lamps item is under this category; then, they backtracked and visited Home Appliances category that contains Lamps item.

As for modelling the data of this study using Chen’s model, this model was developed to model searching a single-level menu not a two-level hierarchical menu as used in this study. Therefore, Chen’s model could not be used to model the data of this study.

The results of this study should be considered in light of some limitations that might affect the results of this study. This study recruited participants through Amazon Mechanical Turk (Mturk). Although recruiting the participants through MTurk allowed conducting this study with a wider and more diverse population within a short time. However, the participants' seriousness is a concern as it might affect the results. It might be that some outlying menu search performance cases were due to participants’ inattentiveness, so they masked the effect of the studied factors. The participants’ inattentiveness was not a concern in Study (2) as it was conducted in a controlled environment that allowed the researcher to observe the participants and see their attentiveness. However, the concern of participants’ inattentiveness in this study was because it was conducted online and no way to observe the participants. Therefore, it would be better for future works that aim to investigate outlying user performance to be conducted in a controlled environment.

Also, this study attempted to link the perception of menu semantics (the semantic organization of menu items) to outlying menu search performance and used card sorting to collect the perception of the semantic relationship between menu items. It might be that menu search performance is not influenced by the perception of menu semantics (the semantic organization of menu items). This was indicated in work by Schmetto and Sommer (2016). Their work tried to link card sorting to browsing performance. Their results suggested that browsing performance is not influenced by having a matching mental model of website structure. They discussed their results and mentioned that menu semantics (category labels) is a key factor in browsing performance and not the well-designed website structure.

Therefore, it might be that outlying menu search performance is due to having a different perception of the menu semantics (category labels). However, the results of Yin (2018) suggested that there is no relationship between having a different perception of menu semantics (category labels) and outlying menu search performance. It might be that the effect of having a different perception of menu semantics was not clear because the menu used in her study contained competitive menu categories that confused most participants. Therefore, future works that aim to investigate the role of having a different perception of menu semantics in outlying menu search performance should focus on the role of the semantic relevancy of menu categories to their corresponding items. Also, they should use a menu with distinguishable menu categories.

Additionally, it was difficult to justify the occurrence of outliers in this study, although several factors were investigated. This might be because of depending on data that are insufficient to identify the reasons behind the outlying menu search performance. Therefore, other methods should be used to collect other data that help in identifying the causes of the outlying performance. This was suggested by Albert and Tullis (2013) that performance data should be supplemented with other data, such as observational or self-reported data, to improve the understanding of the problems and how they might be solved.

Chapter 8

Study 7: Investigating Outlying Menu Search Performance Using the Retrospective Think-Aloud (RTA) Protocol

8.1 Introduction

The findings of Study 6 (Chapter 7) indicate that there was a difficulty in justifying the occurred outlying menu search performance, although several factors were considered and investigated. This difficulty might be because of depending on data that are insufficient to identify the reasons behind the outlying menu search performance. Therefore, other methods should be used to collect other data that help in identifying the causes of the outlying performance. This was suggested by Albert and Tullis (2013) that performance data should be supplemented with other data, such as observational or self-reported data, to improve the understanding of the problems and how they might be solved.

Observation can help in guessing problems from what was seen. However, it cannot help in knowing what users think about it. The think-aloud technique is a useful method to understand what is in the head of users (Preece et al., 2015). The think-aloud protocol (TA) is a widely used method to collect users' thoughts and frustration (Lazar et al., 2017). The TA protocol was first proposed by Ericsson and Simon (1980), and since that time, many researchers and usability testing practitioners have applied this method in usability research and practice. By using TA protocol, the areas that cause users to struggle and the reasons behind this struggle are verbally reported. The usability practitioners utilise this information and combine it with other measures to identify the problematic parts of the application under evaluation and to suggest some improvements accordingly (Olmsted-Hawal et al., 2010). Therefore, the TA protocol was suggested as a method for collecting users' verbal thoughts that might explain why the outlying performance has happened.

Two types of the TA protocol were identified by Ericsson and Simon (1993), concurrent think-aloud (CTA) and retrospective think-aloud (RTA). In CTA, the thoughts are verbalised by the user during performing the task, while in RTA, the thoughts are verbalised by the user after completing the task. Both methods have advantages and disadvantages.

According to Van Den Haak et al. (2003), from a theoretical perspective, using RTA instead of CTA has some advantages and disadvantages. The advantages of using RTA can be specified in four points. First, RTA is less susceptible to reactivity than CTA because participants perform the task in a normal way as usual, and therefore their performance is not changed to be better or worse. The second advantage is allowing measuring the task completion time, which cannot be measured accurately by using CTA which slows down the participant performance as a result of the requirement to verbalise the thoughts while performing the task. The third advantage is the possibility to get more reflections from participants, which can reveal many high-level reasons for individual usability problems. Finally, when usability testing is conducted across cultures and with multiple languages, it is easier for participants to verbalise their thoughts in a foreign language after completing the task than while performing the task.

On the other hand, there are some disadvantages of using RTA instead of CTA, as discussed by Van Den Haak et al. (2003). First, RTA requires a longer testing session with the participants because the participants are not just required to perform the task, but they must watch their performance recording after that. The second disadvantage that should be considered is the possibility that participants produce biased or fabricated thoughts, or they might forget important information that happened while performing the task. However, this mostly relies on the stimuli used to help the participants in recalling their thoughts.

RTA seems more suitable to investigate outlying user performance because of the need to focus on measuring a task completion time which is sensitive to any source of distraction that might be caused by CTA. The CTA is considered an issue in analysing the task completion time data, and the RTA is an appropriate method to collect the most accurate task completion time data (Albert & Tullis, 2013). Moreover, there is a need to focus on the users' verbal thoughts to understand why outlying performance happened. The RTA participants are offered more opportunities to comment on their performance during interacting with the test object (Cotton & Gresty, 2006). The RTA has also shown that it gives more information about users' interpretations and strategies in carrying out tasks (Guan et al., 2006).

This study aimed to investigate the outlying menu search performance using RTA. Therefore, this study involved asking participants to carry out a menu search task, then they immediately watched a screen recording of their performance in the menu search task and reported their thoughts retrospectively.

8.2 Method

8.2.1 Design

The goal of this study was to find out the reasons behind the outlying menu search performance by using RTA. This study was conducted individually. Each participant in this study undertook a menu search task. Then, they directly watched a screen recording of their performance and verbalised what they did and what they thought while doing the task.

The menu search task in this study was almost the same menu search task in Study 6 (Chapter 7), but this study consisted of three menu search trials instead of six menu search trials. The participant's menu search performance (search time) was recorded in each trial. Search time is the time taken by a participant to find and select the target item, and it was measured in seconds. Additionally, a screen recorder was used to record the participant's performance in the menu search task. The screen recording was used then in the RTA session to help participants in recalling their thoughts.

In the RTA session, the participant watched the screen recording of his/her performance in the menu search task and verbalised their thoughts while doing the task. The participant's verbal thoughts were recorded.

8.2.2 Participants

Opportunity sampling was adopted in this study to recruit participants. The announcement of the need for participants has been circulated between the undergraduate students and administrative staff at the Computer and Information Science College (CCIS) at King Saud University (KSU). After that, many students and several administrative staff contacted the researcher to schedule the appointment.

Thirty-three (33) participants were recruited for this study. They were all female as KSU adopts single-gender education. They belonged to two age groups (18-29) and (30-49), see Figure 8.1, 60.6% of participants have good experience in purchasing electronic devices online, 33.3% of the participants have little experience, and 6.1% have no experience, see Figure 8.1.

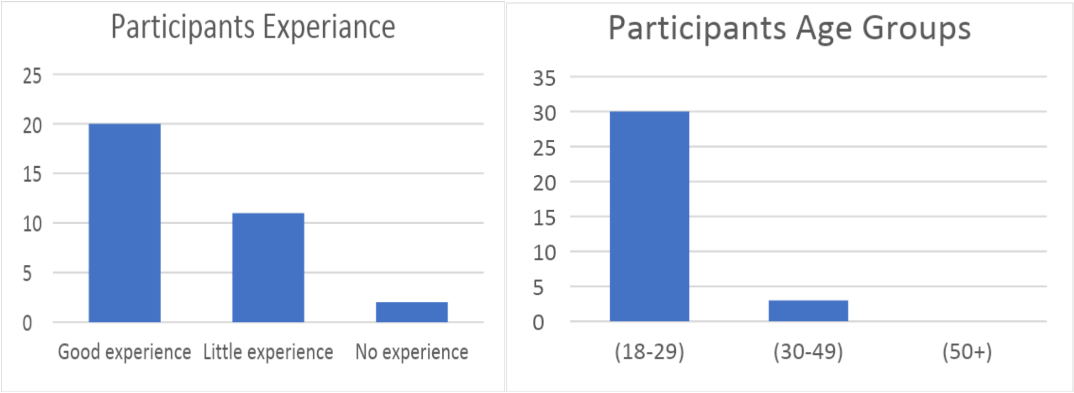


Figure 8.1: Demographics of participants.

8.2.3 Tasks and Materials

It was necessary to make the participant familiar with the testing laptop and with the task as well. Therefore, preliminary tasks were prepared. These preliminary tasks involved an online demographics questionnaire fill-in task and a trial menu search task. The participant should fill in the demographics questionnaire and carry out the trial menu search task before carrying out the main task. The demographics questionnaire included closed questions about age and experience in purchasing electronic devices online. The trial menu search task involved asking the participant to find a specific item (Dairy) in a food menu and select this item as quickly and accurately as they could, see Appendix F.2.

As for the main task (menu search task), the participant was asked to carry out three menu search trials. In each trial, they were asked to search for a specific item in a two-level hierarchical menu and select this item as quickly and accurately as they could. The menu search trial starts once the button “Start task” is clicked and the menu appears on the screen. The menu search trial finishes when participants select the target item.

This study used the same two-level hierarchical menu that was used in Study 6 (Chapter 7). This menu was derived from the electricals menu in the John Lewis online store. The same three menu categories were used (Small appliances, Smart Tech and Home Appliances), and the same items under each category, see Figure 8.2. However, the menu was translated into the Arabic language as all participants were native Arabic speakers. The translated menu was similar to the menus found in the well-known Saudi commercial websites that sell electricals, such as the Extra online store.

Three items (Lamps, Home Telephones and Toasters) were chosen as target items, one target item from each category. In the first trial, the participants were asked to find lamps in the electricals menu. In the second trial, the participants were asked to find home telephones in the electricals menu. In the third trial, the participants were asked to find toasters in the electricals menu. The lamps and home telephones were chosen as target items because they were confusing items, as was found in Study 6 (Chapter 7), as many participants were not able to find these target items directly without backtracking to the main categories. The description of each menu search trial was written clearly and set out at the top of the page, see Figure 8.3.

The menu was developed using the same tool, “Treejack”, that was used in Study 6 (Chapter 7).

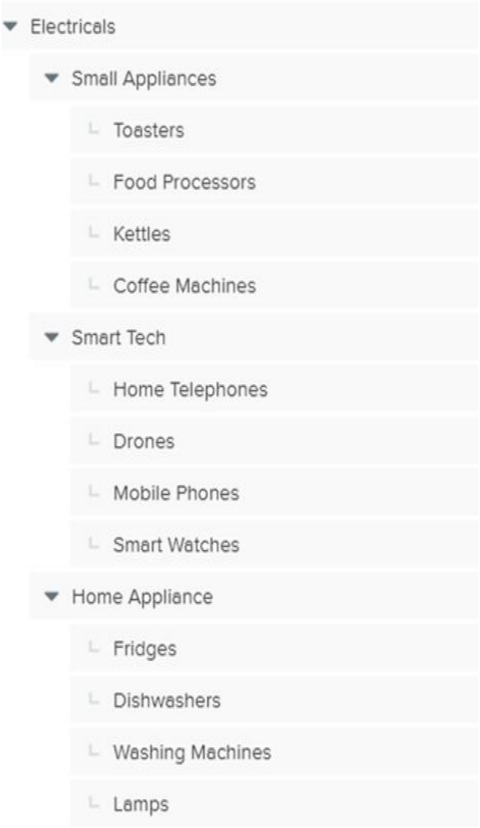


Figure 8.2: The menu used in the main task.

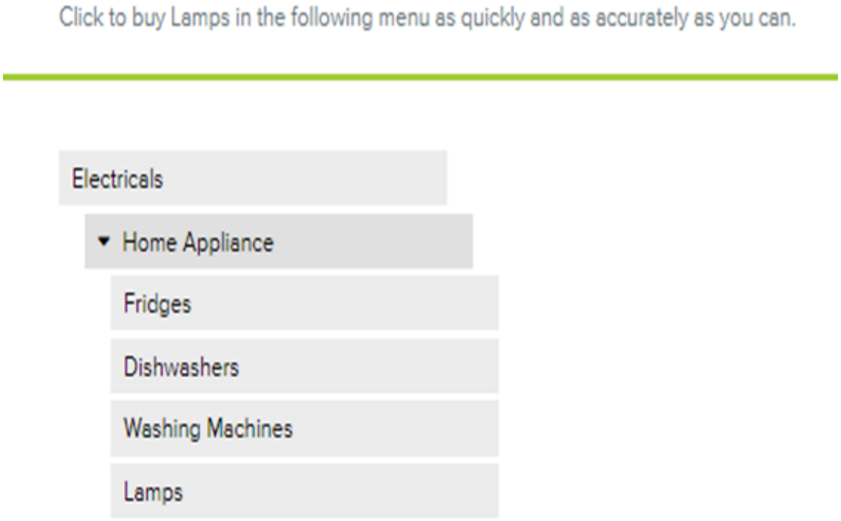

Click to buy Lamps in the following menu as quickly and as accurately as you can.

Figure 8.3: Example of a menu search trial.

8.2.4 Setting and Equipment

All sessions of this study were conducted in one office in the College of Computer and Information Science (CCIS) at KSU. This office consisted of sofas, a comfortable table and chair, a laptop with an external mouse, a bookcase and other items that are usually found in a typical office, see Figure 8.4. The environment and equipment were controlled to reduce the bias that might arise due to participants having different settings and equipment.

To offer a comfortable environment, the sources of noise were avoided, and the light and temperature were set to normal levels. Just the researcher and the participant attended the session. The researcher sat behind the participant to reduce the feeling of being observed.

The same laptop and the same internet connection were used in all sessions. This was to avoid any variation that might affect the user performance measurement. The test laptop was equipped with a 2.40 GHz Intel processor, Windows 10 operating system and an external computer mouse.

To record the screen, Camtasia, a screen recording software, was installed and run on the test laptop. The voice recorder was also run on the test laptop to record the verbal report of the participants.

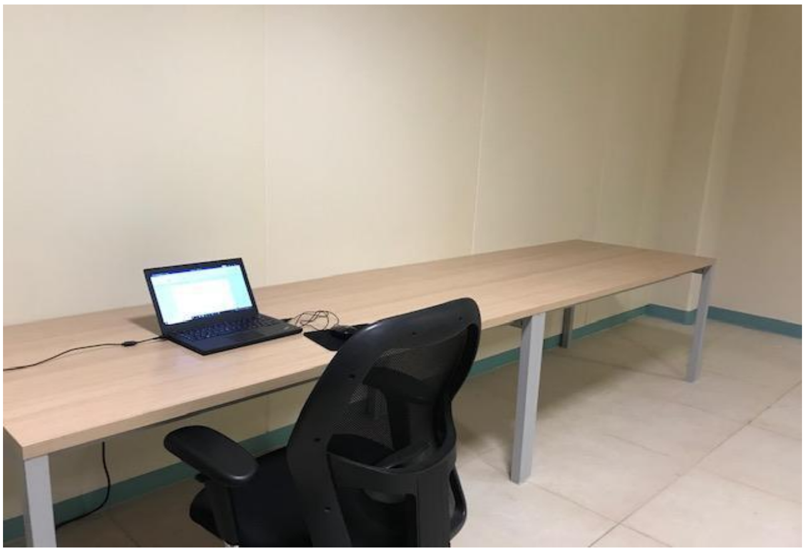


Figure 8.4: The equipment and setting.

8.2.5 Procedure

This study was conducted in 33 individual sessions. When the participant arrived at the office, she was welcomed and made to feel at ease. Then, the participant was asked to read the information sheet that describes the aim of this study and other issues related to the participant's rights, such as withdrawal and confidentiality, see Appendix F.1. After reading the information sheet, she was asked to sign the informed consent form. After that, she was introduced to the study and given the instructions. These instructions were read out from a paper to guarantee consistency. To make the participant familiar with the testing laptop and with the task, she was given a link to the preliminary tasks and asked to open this link and follow the instructions. The link directed the participant to an online demographic questionnaire. After completing the demographic questionnaire, the trial menu search task was presented.

After that, the screen recording software (Camtasia) was set up by the researcher, and the participant was given a link to the main menu search task and asked to carry out the task in silence and without any assistance from the researcher. Once she completed the task, she immediately sat with the researcher to watch the screen recording of her performance in the task, and she was asked to say what was going on in her head during the menu search task. Acknowledgement tokens such as “Ok” and “yeah” were used to show the expected response from the active listener without being directive, as suggested by Boren and Ramey (2000). The retrospective verbalisation was recorded using voice recording software.

At the end of the session, the participant was thanked for her participation and requested not to discuss the study with others who might participate in this study. Following that, the researcher collated all required documents and copied the video and voice recording to a folder identified by the participant ID. After that, the environment was restored to the original situation in preparation for the next session.

8.2.6 Pilot

To find out the most suitable design for this study, a pilot study was conducted to check whether face recording helps in understanding participants' interaction behaviour. Also, to check whether doing multiple trials influences the participant's ability to recall the thoughts. Therefore, four conditions were tested: face recording and single trial, face recording and multiple trials, no face recording and single-trial and no face recording and multiple trials.

Four groups of participants took part in this pilot study, two participants in each group. The first group were asked to complete a single menu search trial. Their performance on the task and their faces were recorded using Camtasia. After completing the task, each participant watched the video recording of her performance and was asked to verbalise her thoughts. The second group did the same as the first group, but multiple menu search trials should be completed by this group. The third group did the same as the first group, but with no face recording. The fourth group did the same as the second group, but with no face recording.

It was noticed in this pilot study that face recording concerned the participants as they seemed hesitant to accept that. Additionally, the face recording did not help in understanding the user’s behaviour as most users showed a masked face. Therefore, it was decided that participants’ faces would not be recorded.

Regarding doing multiple menu search trials, it was found that doing three menu search trials did not affect the ability of the participants to recall their thoughts in the earlier trial. This might be because each trial was simple and similar to the other trials. Therefore, it was decided to set three menu search trials in the main task.

No major issues were raised while piloting this study, just minor issues such as the font size of the displayed text. This issue was solved by zooming up the web browser.

8.2.7 Data Analysis

Once the 33 sessions were completed, the collected data were prepared for analysis. Three types of data were collected: menu search performance data (search time, path), screen recording and RTA verbal reports.

The menu search performance data for each menu search trial was recorded by the Treejack tool. These data were downloaded as an excel sheet. Then, the boxplots were used to display the distribution of the search time and identify outliers in each menu search trial.

The RTA verbal thoughts recordings were transcribed into text files. After that, the transcription was analysed using thematic analysis. The thematic analysis was the most suitable qualitative analysis method because this study has a focused aim which is in finding out the causes of outlying menu search performance.

To identify the themes in this study, the same approaches used in Study 1 (Chapter 3) was also used in this study. The inductive ‘bottom-up’ approach, the semantic approach and the essentialist/realist analysis approach. Also, Braun and Clarke’s six-phase guide of the thematic analysis was followed. First, the transcription was read and re-read to get familiar with the content. Then, each segment of the transcription that was relevant to the menu search performance was identified and coded. Open coding was used, which means that codes are developed and changed through the coding process. Each piece of data was given a labelled code that was added to the transcriptions document as a comment beside the highlighted interesting data. Examples of the generated codes were strategy and difficulty.

Then, all the generated codes were grouped and collated to search for potential themes. The similarities and differences among the generated code were identified – this helped in making sure that the generated codes and their corresponding data extracts could form themes. The two identified themes were: menu search strategies and menu search obstacles.

After that, the two identified themes were reviewed to make sure that, for both themes, there were enough data to support them and that there is consistency in the data that belonged to the same theme. Moreover, the identified themes were reviewed to make sure that they reflect all interesting issues in the entire collected data.

Once satisfied with the identified themes, each theme was defined and named for the analysis. Producing the report phase was done concurrently with the defining and naming themes phase because defining and naming themes phase necessitates writing about each theme along with the corresponding data extracts, see the Results section. The two themes were reported in a narrative way that ultimately presented the findings.

To identify the reasons behind the occurred outlying menu search performance, each outlying performance case was investigated in detail. The outliers’ verbal thoughts transcriptions, along with their screen recordings, were analysed thoroughly. It was noticed that one outlier did not follow the instruction that emphasised reading the task description before starting the task. She pressed the ‘Start task’ button. Then she read the task description. This, of course, affected the search time and made it much longer.

Therefore, the performance data of this participant in this task was removed from the analysis.

8.3 Results

8.3.1 Menu Search Performance (quantitative results)

The average search time for each menu search trial was presented in Table 8.1.

Trial #	Mean (SD) search time in second	Percentage of direct success
1: select Lamps	12.13 (5.96)	41%
2: select Home Telephones	7.63 (3.39)	65%
3: select Toasters	10.10 (3.90)	18%

Table 8.1: The average menu search performance in each menu search trial.

The average search time in trial 1 (select Lamps) and trial 3 (select Toasters) was longer than in trial 2 (select Home Telephones), see Table 8.1. This was because many participants in trial 1 and trial 3 could not find the target item directly, as indicated by the percentages of direct success (choosing the target item without backtracking) for each trial, see Table 8.1. This means that many participants in trial 1 and trial 3 first visited the wrong category, and then backtracked to find the right category. Therefore, they spent a long time finding the target.

As for the outlying menu search performance, outliers were found in each trial, see Figure 8.4. The total number of the outlying menu search performance cases was 3 cases created by 3 participants, 2 cases in trial 1, and one case in trial 2

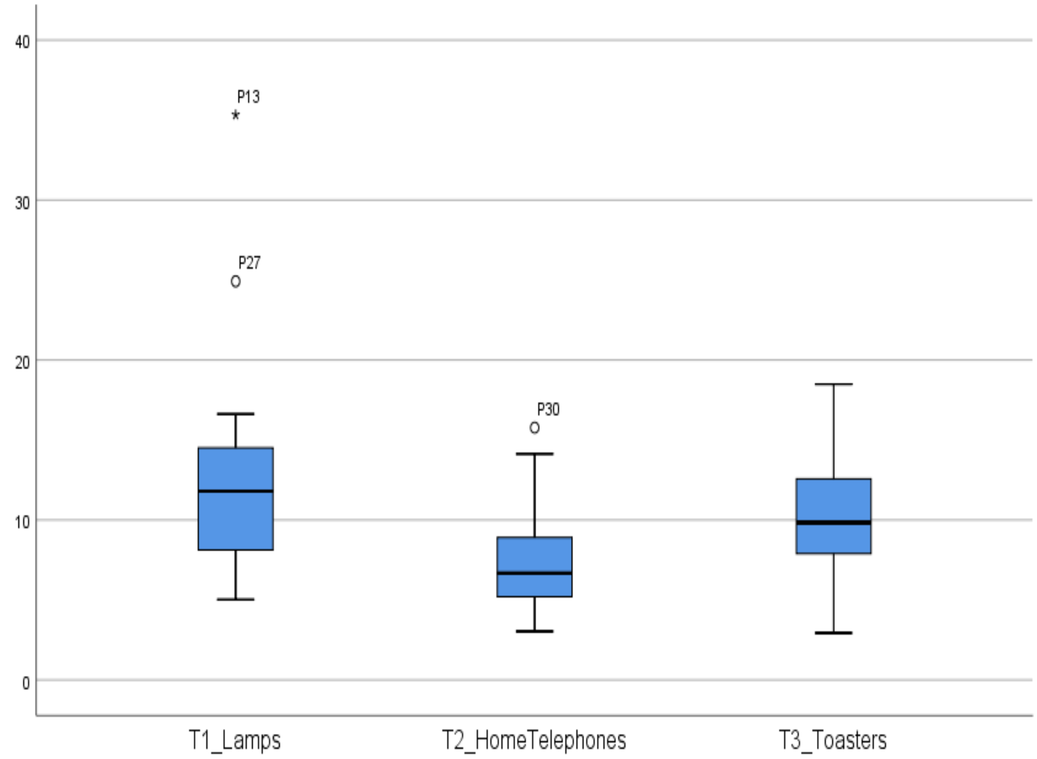


Figure 8.5: Boxplots of menu search time in the three trials; trial 1 (select Lamps), trial 2 (select Home Telephones) and trial 3 (select Toasters).

8.3.2 Menu Search Performance (qualitative results)

By analysing the collected verbal reports from the participants, two main themes were found related to the menu search performance. These two themes were presented in the following sections.

- **Menu Search Strategies**

The most common strategy used by the participants to search for a specific item in a two-level hierarchical menu was satisficing strategy. They selected the good enough category that might lead to the target item based on the semantic relevancy of the category label to the target item. The following data extracts explain these strategies:

> *"To search for lamps, I thought that lamps related to home appliances. Then, I found it directly" P(11);*

> *"I expected to find lamps under small appliances because they are small. After that, I found it under home appliances" P(15);*

> *"Searching for home telephones was clear for me because I knew it was under smart devices from a smart word." P(23).*

Another strategy that was adopted by some participants is optimizing. They evaluated all the categories to find the optimal category, as mentioned by the participants in the following extracts:

> *"When I started this task, I checked the displayed categories to decide where lamps can be found, I was confused whether lamps are under home appliances category or small appliances category" P(13);*

> *"As this was the first time I saw this menu, I tried to read the categories one by one to figure out where I can find the lamps. I was confused between small appliances and home appliances. I tried small appliances and then home appliances" P(27).*

After interacting with the menu, some participants learned this menu and were able to find the target item directly, as stated by some participants in the following extracts:

> *"Finding home telephones was straightforward for me because I already saw this list" P(1);*

> *"In the third trial, I saw toasters before. It was under small appliances. Therefore, I found it immediately" P(10);*

> *"Finding toasters was easy because I saw it before. Therefore, I chose it immediately" P(13).*

- **Menu Search Obstacles**

The participants faced some obstacles while carrying out the menu search. One of these obstacles was uncertainty. The participants were confused when they were asked to search for a specific item, and they expected that more than one category could lead to the target item (competitive menu categories). They faced difficulty in deciding which was the best one that could lead to the target item. This problem was very noticeable in the first trial which was about searching for lamps. This might be because lamps can be classified under small appliances or home appliances. Therefore, it confused many participants, as stated in the following data extracts:

> *"I was really confused about selecting small appliances or home appliances. It was not clear for me" P(1);*
>
> *"Here, when I was asked to search for lamps in the electricals menu, I was confused about small appliances or home appliances. Then, I found it under home appliances." P(28);*
>
> *"To find lamps, I was confused about whether to be under small appliances or home appliances. I felt that I do not know which one, then I selected small appliances as lamps are small. I did not find it. Therefore, I said it is definitely under home appliances, and I found it." P(31).*

This uncertainty was also indicated in the screen recording when participants moved their mouse cursors up and down over the menu categories.

Another menu search obstacle was having a different perception of menu semantics (the semantic relevancy of a menu category to its corresponding items). Some participants had a completely different perception. Therefore, they visited several wrong categories until finding the target item. As outlined by the participants in the following data extracts:

> *"I went to small appliances because home telephones are small devices when I did not find it, I went to home appliances because telephones are related to homes, then I went to smart devices after several attempts" P(30);*
>
> *"It was required to find toasters. I expected that it is under home appliances, but I did not find it. Then, I expected that is under smart devices, again I did not find it. Finally, I found it under small appliances although I did not expect that" P(6).*

8.3.3 The possible causes of the occurred outlying menu search performance

By investigating the two outlying menu search performance cases in the first menu search trial (select Lamps), it was found that both outliers P13 and P27, were confused and uncertain due to competitive menu categories. This was indicated by the participants' verbal thoughts as stated in the following data extracts:

> *"When I started this task, I tried to find the category that lamps belong to. I was confused between small appliances and home appliances, I clicked on small appliances, and I did not find it. Then, I stayed confused and then went to the home appliances" P(13);*

> *"As this was the first time I saw this menu, I tried to read the categories one by one to figure out where I can find the lamps. I was confused between small appliances and home appliances. I tried small appliances and then home appliances" P(27).*

Their uncertainty was also indicated in the screen recording of their performance. Their mouse cursor was moving up and down over the menu categories. Additionally, I noticed that during the RTA session, they seemed frustrated due to the uncertainty caused by the competitive menu categories.

Although several participants suffered from uncertainty in the first trial (select Lamps), as mentioned in the previous section, the outliers P13 and P27 were terribly affected.

By investigating the third outlying menu search performance case, which was in the second menu search trial (select Home telephones), it was found that outlier P30 was struggling in finding the item (Home Telephones). Her struggle was because she had a completely different perception of menu semantics, as she stated in the following data extracts:

> *"I went to small appliances because telephones are small devices when I did not find it, I went to home appliances because telephones related to homes, then I went to smart devices after several attempts" P(30).*

This was also indicated by the several wrong paths she tried to find the item, although this trial was straightforward for most participants.

8.4 Discussion

This study aimed to investigate the outlying menu search performance using the RTA. This study consisted of a menu search task, followed by an RTA session in which participants were shown a screen recording of their performance in the menu search task and asked to verbalise their thoughts retrospectively.

The overall results showed that were several outlying menu search performance cases in the menu search task, and the RTA helped in identifying the causes of these cases. Two causes were identified. The first one was the uncertainty caused by the competitive menu categories. Although several participants faced uncertainty due to the competitive menu categories, outliers struggled the most and spent a long time completing the task compared with the others. They were terribly affected. This means that participants were different in tolerating uncertainty caused by the competitive menu categories. Intolerance of uncertainty is an individual difference variable that is associated with neuroticism personality (Fergus & Rowatt, 2014). These results indicate that individual differences in personality may play a role in the outlying user performance. This seems to be inconsistent with the findings of Study (6) where there was no evidence that outlying performance is due to specific participants as the occurred outliers did not show outlying performance on a regular basis. This might be because specific individual differences might be highly correlated with specific design features or specific tasks (Dillon & Watson, 1996). Therefore, the effects of these individual differences might not be persistent. The outlying menu search performance in this study was understandably caused by the interaction effect between the participant's personality (intolerance of uncertainty) and competitive menu categories. The intolerance of uncertainty was triggered by the competitive menu categories.

The second cause of outlying performance in this study was having a completely different perception of menu semantics. One outlier had a different perception of menu semantics. Therefore, she struggled to find the target, although most participants found the target straightforward. This finding agrees with the prediction of Schiller and Cairns (2008)' model. Their model predicted outliers as an outcome of variation in perceiving the menu semantics.

The collected verbal reports from the participants helped in understanding the menu search strategies that were followed by participants to search for a specific target item in a two-level hierarchical menu. The most common strategy used by the participants was satisficing strategy. They selected the good enough category that might lead to the target item based on the semantic relevancy of the category label to the target item. This strategy was described by Simon (1955). He stated that people do not look for the best solutions to problems rather they solve the problem within bounded rationality where time and cognitive limitations prevent them from evaluating all possible solutions. Another strategy adopted by the participants was optimizing strategy. They evaluated all the categories to find the optimal category.

Also, the collected verbal reports helped in understanding the obstacles that faced the participant while searching a two-level hierarchical menu. The first one was the uncertainty that made the participants face difficulty to determine the optimal category that will lead to the target item. This uncertainty was due to the competitiveness of two categories or more, which made the participants confused. This confusion was also apparent in the screen recording when the participants moved the mouse cursor up and down over the categories. This finding is consistent with the findings of previous works that studied the menu search behaviour, such as Brumby and Howes (2003) and Cox and Young (2004). Their works highlighted the interdependency between the assessment of each menu item and how this affects the menu search behaviour. Cox and Young (2004) demonstrated that in the exploration of menus, the existence of distracter items that received a similar scent assessment to the target item (mediocre distracters) would result in difficulty in identifying the best item among these distracter items. Therefore, to decide which item will lead to the goal, additional cognitive efforts are required. As a result, multi-scanning behaviour on this menu is more likely to be shown.

This obstacle highlighted the importance of choosing proper labels to name the menu categories. The label of a category should reflect the corresponding items while not overlapping with other labels (Bailly et al., 2016). The label should be distinguishable to avoid being confused with another label (Preece et al., 2015).

The second identified obstacle was having a different perception of menu semantics. The participants who have a very different perception of menu semantics struggled to find the target item. This obstacle was stated by Norman (2013), who noted that the difference between the user mental model and designer mental model is the main source of many problems in interaction with technologies.

In summary, menu semantics is an influential factor in menu search performance. While the menu semantics helps in searching a menu, it can simultaneously hinder searching a menu. Therefore, the menu semantics may play a role in outlying menu search performance if it is perceived very differently by some individuals. Additionally, menu semantics may play a partial role in outlying menu search performance when semantically competitive menu categories terribly affect some individuals who might be intolerant of uncertainty.

In this study, it was possible to unconfound the effect of menu semantics and the effect of a slow user on performance time. The outlier was due to the perception of menu semantics occurred as an outlier only in the second menu search trial. This indicates that her outlying performance was not due to the slowness as she was not slow (outlier) in the other menu search trials. Additionally, it was clear that she had a completely different perception of menu semantics as indicated by her verbal report. She tried several wrong paths to find the target item, although this trial was straightforward for most participants. This indicates that her outlying performance was due to her different perception of menu semantics.

The number of menu search trials in this study was few. However, these few trials were enough as a sample that represents menu search tasks involving finding and selecting a specific target item in a two-level hierarchical menu.

The results of this study should be considered in light of some limitations that might affect the validity of these results. The results suggested that outlying menu search performance can be caused by the interaction effect between the participant’s personality (intolerance of uncertainty) and competitive menu categories. However, these are preliminary results and not conclusive. It was just based on an analysis of outliers’ performance and verbal reports supported by a researcher's observation of the outlier’s feelings during the RTA session. To confirm the role of participants’ personalities in the outlying menu search performance, further work can be done. This work should measure the participant's personality using the big five model, also called the OCEAN model (openness, conscientiousness, extraversion, agreeableness, neuroticism). Then, it should analyse the relationship between personality and outlying menu search performance.

Chapter 9 Conclusion

9.1 Introduction

Outlying performance is a potentially important feature of usability tests. There is always one or more participants who take a long time to complete a task compared to other participants. Those participants appear as outliers in the collected time-on-task data. Such outliers can skew any statistical analysis, but, at the same time, they may indicate genuine problems with the interaction.

This research is about outlying performance in menu search. It investigated the role of the perceived menu semantics in outlying menu search performance. In addition, it checked whether outlying menu search performance is due to specific individuals.

In this final chapter, an overview of the research is provided. Additionally, the contributions of this research and the implications of these contributions are presented. In addition, the limitations of this research and the opportunities for future work are discussed.

9.2 Research Overview

Three main research questions have been addressed in this research. The first research question was "How are outliers interpreted and treated in usability testing practice?". The second research question was "Does the perceived menu semantics play a role in outlying performance in menu search?". The third research question was "Does outlying performance in menu search occur due to specific individuals?".

To answer the first research question, Study 1 (Chapter 3) investigated outlying performance in usability testing practice. Study 1 aimed to gain knowledge of how practitioners interpret and treat the outlying performance incidents in usability tests. This study found that the interviewed usability practitioners seem aware of the regular occurrence of outliers in usability testing. They tend to link outlying performance cases to individual differences instead of linking them to usability problems, and there is no systematic approach to addressing them.

The focus of this research then moved to answering the second research question. Several studies have been conducted to investigate the role of perceived menu semantics in outlying performance in menu search. Study 2 (Chapter 4) investigated whether poor semantic organization of menu items plays a role in outlying menu search performance. The findings of this study suggest that menu semantics may play a role in this respect as more outlying performance cases were found in searching a poorly organised menu. However, it could be that other factors caused these outlying menu search performances. Therefore, more studies using different methods are needed to check the role of menu semantics in outlying menu search performance.

Therefore, a modelling approach was adopted to help explain the role of menu semantics in outlying menu search performance. As the adopted menu search model is based on ML, it needs to be trained on menu samples to learn menu search strategies. Constructing menu samples needs collecting semantic similarity ratings of menu items from human participants. Therefore, Study 3 and Study 4 (Chapter 5) were conducted to collect the semantic similarity ratings of menu items. Study 3 used pairwise similarity ratings to collect the semantic similarity ratings from the participants. Study 4 used card sorting to collect the semantic similarity ratings from the participants. Study 5 (Chapter 6) was then conducted to check whether the adopted menu search model can predict outlying menu search performance due to menu semantics. The model predicted outliers, but they were not due to menu semantics. Therefore, this study found no evidence that outlying performance is related to menu semantics.

Following that, Study 6 (Chapter 7) was conducted to investigate whether outlying menu search performance is due to having a different perception of menu semantics. In addition, this study tried to answer the third research question by investigating whether outlying performance is due to specific individuals. This study found no evidence that the outlying menu search performance is related to having a different perception of menu semantics. In addition, no evidence was found that the outlying menu search performance is due to specific individuals. It was difficult to justify the occurring outlying menu search performance cases in this study. Therefore, it was suggested that there is a need to collect other data such as self-reported data that might help in identifying the causes of outlying performance in menu search.

Consequently, Study 7 (Chapter 8) was conducted to investigate the outlying menu search performance using a retrospective think-aloud protocol. This study found two possible causes of outlying menu search performance. The first one is related to individual differences in tolerating the uncertainty caused by the competitive menu categories. The second cause was having a different perception of menu semantics.

9.3 Research Contributions

To the best of my knowledge, only three studies, discussed in Chapter 2 (section 2.4.3), investigated the outlying performance in menu search and reported preliminary results. The series of empirical studies reported in this thesis contributed to our general understanding of outlying performance and why it might happen in menu search. The contributions of this research, along with their implications, are discussed in the following sections.

- **There is always one outlier**

According to Schiller and Cairns (2008), in usability tests, there is regularly one participant who is noticeably slower than all others. All studies conducted in this thesis and collected time-on-task data confirmed Schiller and Cairns' observation. There is always one or more outlier(s).

The regular occurrence of outliers in usability tests indicates that outlying performance is an important feature that needs to be addressed systematically. Unfortunately, the literature does not help in understanding why outliers regularly occur, and accordingly how to handle them. The absence of guidance is critical as this might affect the interpretation of the outcomes of usability tests.

- **The interviewed practitioners seem aware of outliers in usability testing, but there is no systematic approach to addressing them**

Several interviewed practitioners explicitly mentioned that they always find outliers. They tended to link outlying performance to individual differences instead of to usability problems. They claimed that they always consider outliers and they do not exclude them unless they have a strong justification for doing so. They suggested different *post-hoc* strategies for treating outlying user performance cases. However, no systematic approach was followed by the interviewed practitioners when it came to dealing with outliers in usability tests. Therefore, several suggestions were put forward for practitioners when it came to handling outliers in usability tests (see section 9.4).

The practitioners' awareness of regular occurrences of outlying user performance also confirms the observation of Schiller and Cairns (2008) that there is always one outlier in usability tests. Additionally, the lack of a systematic approach to treating outliers in usability testing practice might reflect the lack of work in the literature that investigates the causes of this problem and how to treat it accordingly.

- **Perceived menu semantics may play a role in outlying menu search performance**

This research focused on outlying performance in menu search. It was motivated by Schiller and Cairns (2008) work that identified outliers in menu search and, as a result of modelling work, attributed this to the perceived menu semantics. However, modelling alone is not an alternative for gathering and analysing data on real users searching menus to provide empirical evidence that backs any claim. Therefore, this research addressed this limitation and empirically investigated the role of perceived menu semantics in outlying menu search performance.

The last study in this research found that menu semantics may play a role in outlying menu search performance if it is perceived very differently by some individuals. This finding agrees with the prediction of Schiller and Cairns (2008)' model. Their model predicted outliers as an outcome of variation when it comes to perceiving menu semantics. However, some studies in this research found no evidence that perceived menu semantics plays a role in outlying menu search performance as the occurred outlying performance cases were not related to the perceived menu semantics. Therefore, perceived menu semantics may be a factor in outlying menu search performance, but it is not the only factor.

This finding implies that outlying performance in usability testing should be considered as it might represent a problem in interaction with a tested computer interface, at least for some users. In addition, it implies that outlying performance can be fixed by making the design better for outliers. For example, menu semantics can be improved by offering several correct paths that reflect a variety of menu semantics perceptions.

- **Outlying menu search performance can be due to individual differences**

One of the possible causes of outlying menu search performance that was found in this research is the intolerance of uncertainty that is caused by competitive menu categories. The outliers struggled and spent a relatively long time completing the task due to a degree of uncertainty. Intolerance of uncertainty is an individual difference variable that is associated with neuroticism personality (Fergus & Rowatt, 2014). This indicates that individual differences in personality can play a role in outlying menu search performance. However, the findings of other studies in this research suggest that there is no evidence that outlying menu search performance is due to specific individuals as the outliers did not keep occurring in all or most menu search trials, rather their performance varied from one menu search trial to another. These inconsistent findings might be because specific individual differences might be highly correlated with specific design features or specific tasks (Dillon & Watson, 1996). Therefore, the effects of these individual differences might not be persistent.

This finding is consistent with the tendency among the interviewed practitioners to link outlying performance to individual differences such as personality.

This finding has an implication for interface design. The interface designers should consider individual differences in tolerating uncertainty. Therefore, they should avoid aspects that cause uncertainty for the users. For example, menu designers should offer differentiated menu categories to reduce the uncertainty caused by competitive menu categories. This will help in improving the total user experience.

Additionally, this finding has an implication for interpreting results in usability testing. The researchers and practitioners should pay attention to possible variances in participants' performance due to the influence of their personalities. In the same way that the influence of age and cultural background are examined, personality must also be examined when interpreting the results. This can be done by the acquisition of personality data by integrating a personality questionnaire with a demographic questionnaire (Schmidt et al., 2019).

9.4 Suggestions for Usability Testing Practitioners

To deal with outliers in usability testing practice, several suggestions are proposed for practitioners:

1. Outliers regularly occur in usability tests. Therefore, always check your data to see if there are outliers.

All conducted studies in this research found that there is always one or more outlier. Therefore, there is a need to check any collected time-on-task data to see if there are outliers.

2. Outliers can be valid data that should not be removed but rather analysed with all other data.

The conducted studies in this research found no simple reason for the occurrence of outliers in usability tests. Outliers were not just due to menu semantics, nor were they due to specific individuals. Therefore, outliers should not be simply removed as this might lead to removing an important insight. They should be considered and analysed with all other data.

3. During data analysis, you need to be aware of the effect of outliers on the whole data. Therefore, robust statistical tests that are resistant to outliers ought to be considered.

The conducted studies in this research found that outliers regularly occur. Such a regular occurrence in the data leads to the distribution does not fit with a normal distribution. Therefore, nonparametric statistical tests can be considered when this is the case. These nonparametric tests are robust to outliers.

Given these above-mentioned suggestions, it might be also useful to develop a plan to deal with outliers ahead of gathering any data. This plan should include strategies, information resources, related factors, and potential results. This plan will help you in deciding on accepting or rejecting outliers as a usability problem if you find outliers.

9.5 Research Limitations and Future Work

The research contributions discussed in this research need to be considered in light of the limitations of the studies conducted in this research. Therefore, further work is suggested to address these limitations. Also, work is suggested to develop the topic in future directions.

The first limitation was getting null results. Study 6 (Chapter 7) found null results regarding whether outlying menu search performance is due to specific individuals. These null results might be because of the small sample size (93 participants) or because the number of tasks was few (6 tasks). Therefore, future works that aim to investigate whether outlying user performance is due to specific individuals need to have a bigger sample size and a larger number of tasks.

Another important limitation of this research is that there are no obvious statistical methods that can be used to compare the number of outliers between different distributions. For example, Study 2 (Chapter 4) could not use Chi-Square to test the relationship between menu organization and outlying menu search performance. This was because one assumption of this test was not met, which is that the expected values should be at least 5 and as outliers are very low in a dataset, therefore, the expected values are less than 5. This issue might be solved by having a bigger sample size. However, the extreme level of outliers is also important, not just their number. Therefore, there is a need to consider this situation and develop new statistical methods and validate them through simulation.

Additionally, this research solely focused on investigating outlying user performance in interaction with menus and, more specifically, linear menus where items are organised vertically. It would be interesting in future works to investigate outlying user performance in interaction with other types of menus, such as mega menus where items are displayed in a two-dimensional layout. Additionally, future work could investigate outlying user performance on other important interactions such as browsing a website and finding a function on a smartphone.

In addition, the menu search environment used in this research was not realistic. The menus were simplified and presented alone on a website page, and there were no surrounding features and sections as are usually found in actual commercial websites.

Using such a simple context that was free from any kind of distractions was done to control the factors that might affect participants' performance. However, it would be interesting for future works to investigate outlying user performance in real-world settings.

Additionally, all studies with the exception of Study 6 (Chapter 7), recruited female participants. This is because I conducted these studies at King Saud University (KSU) which adopted a single-gender education model. It would be interesting for future works to recruit participants from both genders and check whether gender difference plays a role in outlying user performance in usability tests.

Also, this research solely focused on one type of outlier in usability tests - a long task completion time. This is because outliers always occur in this performance metric (Albert et al., 2010; Schiller & Cairns, 2008). However, it would be interesting for future research to look at other types of outliers, such as repeated errors and rating scale points which occur outside the range of other participants' ratings.

The findings of this research suggest that a participant's personality may play a role in outlying performance. To investigate the role of participants' personalities in the outlying performance, further work could be done. This work should measure the participant's personality using the big five model, also called the OCEAN model (openness, conscientiousness, extraversion, agreeableness, neuroticism). It should then analyse the relationship between personality and outlying performance.

This research provides evidence that outliers occur regularly. However, no systematic approach is used by practitioners to treat outliers in usability tests. Therefore, there is a need for future work that proposes and validates guidelines for handling outlying performance in usability studies. These guidelines might include a thorough systematic approach and related issues to deal with outliers. It is important to have such guidelines as it influences the interpretation of usability test results.

9.6 Concluding Thoughts

This research aimed to investigate outlying performance in menu search. In fact, investigating outlying performance was not straightforward. It needs to have a combination of data that supplement each other and help determine the reasons behind the outlying performance.

This research implies that outliers are a fact that should not be ignored. The reasons behind them should be identified. And based on the identified reasons, there is a possibility to improve the design for the outliers. Such improvements may have no significant impact on the average user performance. However, the performance of the outliers could be considerably improved by these design improvements. Consequently, there is a chance to help the outliers and improve the quality of their interaction with systems. This would be significantly considered one of the goals of universal usability that was introduced by Shneiderman (2000).

10. Appendices

Appendix A

Chapter 3

A.1 Information Sheet and Informed Consent for Study 1

Overview

In this study, you will be asked to answer interview questions. The interview questions will be about your experience as a usability practitioner in conducting usability tests and dealing with user data that is noticeably different from the other user data. The interview might take about 40 – 50 minutes.

Questions

If you have any questions about the study, please feel free to ask.

Withdrawing

You have the right to withdraw from the study at any time without giving any justifications.

Data

The collected data (your answers) will be visible to me and my supervisor Dr Paul Cairns. This data will be in the form of a text, but you are not individually recognisable from the data. The data may be involved in publication. However, in the case of publishing this work, you will not be recognisable in any way.

Participant consent

Your participation in this study is entirely voluntary; there will be no remuneration for the time you spend answering the interview questions. All data gathered from the interview will be treated in a confidential fashion. It will be archived in a secure location and interpreted only for this study's purposes. When your data are reported or described, all identifying information will be removed. There are no known risks to participation in this study, and you may withdraw at any point. Please feel free to ask the researcher if you have any other questions; otherwise, if you are willing to participate, please sign this consent form.

Participant's Signature:

Researcher's contact details:

Researcher: Hend Albassam (haa522@york.ac.uk).

Supervisor: Dr Paul Cairns (p.cairns@yor.ac.uk).

A.2 Demographics of Participants

ID	Age	Gender	Years of Experience	Job Role	Organization Type	Corporate UX Maturity
P1	24	Female	1 year	Product Development Specialists \ UX Specialist	Software Company	3: Skunkworks
P2	25	Male	2 years	Software Engineer	IT Solutions Company	4: Dedicated Budget
P3	27	Female	2 years	UX Researcher & Product Designer	Bank	5: Managed
P4	26	Female	4 years	UX/UI Specialist	Technologies Company	4: Dedicated Budget
P5	25	Female	3 years	Senior UX Researcher	Usability Lab	7: Integrated User-Centred Design

P6	29	Male	7 years	Senior UX Designer/ Design Sprint Facilitator	Business Services Company	5: Managed
P7	34	Male	5 years	Mobile UX Researcher	IT Solutions Company	4: Dedicated Budget
P8	25	Female	2 years	UX Specialist	Business Services Company	5: Managed
P9	28	Male	3 years	UX Specialist	Business Services Company	5: Managed
P10	40	Male	14 years	UX Researcher	Innovation and Design Consultancy Company	7: Integrated User-Centred Design
P11	31	Female	3 Years	UX Designer &Analyst	Health Insurance Company	3: Skunkworks
P12	29	Male	6 years	Associate Manager, Research	Innovation and Design Consultancy Company	7: Integrated User-

				& Insights.		Centred Design
P13	25	Male	2 years	UX Consultant	Bank	4: Dedicated Budget

A.3 Nielsen’s Usability Maturity Model

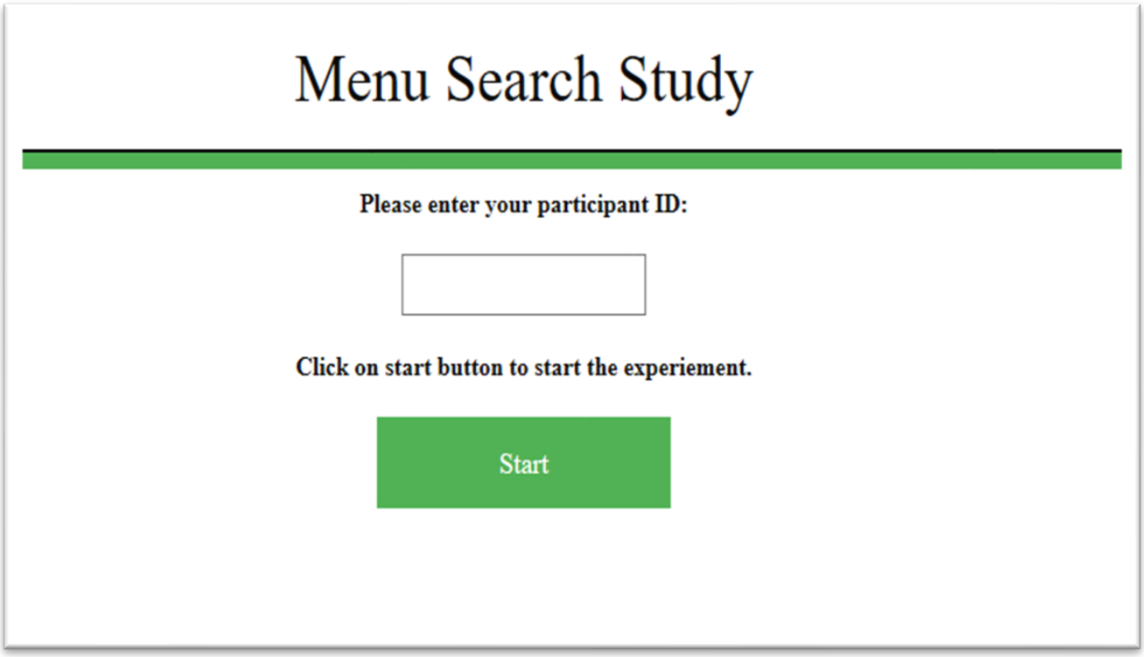
UX Maturity Stage	Featuring	Time to next stage
1: Hostility	Developers simply don’t want to hear about users or their needs	Up to decades
2: Developer-centred	Design team relies on its intuition	2-3 years
3: Skunkworks	Guerilla user research or external usability experts	2-3 years
4: Dedicated Budget	Usability is planned for	6-7 years
5: Managed	Someone to think about usability across the organisation	6-7 years
6: Systematic Process	Tracking user experience quality	6-7 years
7: Integrated User-Centred Design	Employing usability data to determine what the company should build	~20 years
8: User-Driven Corporation	Usability affects corporate strategy and activities beyond interface design	~40 years to get from start

(source: https://www.slideshare.net/Hienadz.Drahun/ux-maturity-models)

Appendix B

Chapter 4

B.1 Material

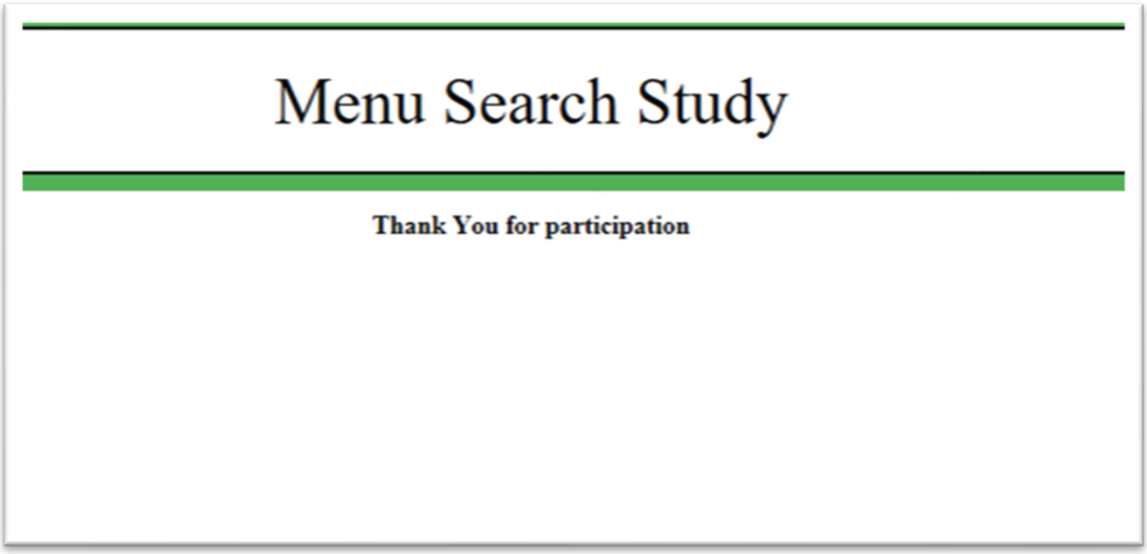


Menu Search Study

Please enter your participant ID:

Click on start button to start the experiement.

Welcome Page



Menu Search Study

Thank You for participation

Thank Page

B.2 Information Sheet and Informed Consent for Study 2

Overview

In this study, you will be asked to search for a specific target in a menu and select this target as quickly and accurately as possible. I will gather information on how you find a target in a menu. In addition, I will gather some demographic data about you. This study will take about 5-10 minutes.

Questions

If you have any questions about the study, please feel free to ask. However, when the study has started, kindly leave your questions until the end.

Withdrawing

You have the right to withdraw from the study at any time without giving any justifications, and do not worry there will be no negative consequences affecting you as a result of your withdrawal.

Data

The collected data will be visible to me and my supervisor Dr Paul Cairns. This data will be in the form of a spreadsheet, but you not be individually recognisable from the data. The data may be involved in publication. However, in the case of publishing this work, you will not be recognisable in any way.

Participant consent

Your participation in this experiment is entirely voluntary; there will be no remuneration for the time you spend in this experiment. However, your participation is highly appreciated as it will help me to understand how people find a specific target in a menu, and this is important in my research which aims at understanding the factors that affect user performance during menu search.

If you are aware of any medical or other condition that might make it unsafe for you to participate in this experiment, please do not sign this form or discuss your concerns with the experimenter before doing this experiment. If you are willing to participate, please sign this consent form and proceed with the experiment.

Participant's Signature:

Researcher's contact details:

Researcher: Hend Albassam (haa522@york.ac.uk).

Supervisor: Dr Paul Cairns (p.cairns@yor.ac.uk).

Appendix C

Chapter 5

C.1 Information Sheet and Informed Consent for Study 3

Overview

In this study, you will be asked to rate the semantic similarity of pairs of menu items. You will be given a list of pairs of menu items, and you must rate them according to what you think about how close in meaning the two items in each pair are.

Examples: (Mountain - Valley) and (Uncle - Brother) are examples of items that are close in meaning. (Apple - Diamond) and (Cat - Chair) are examples of items that are not close in meaning.

There are no right or wrong answers. We are just interested in how similar you find different items. In addition, we will gather some demographic data about you. This study will take about 20-30 minutes.

Questions

If you have any questions about the study, please feel free to ask. However, when the study has started, kindly leave your questions until the end.

Withdrawing

You have the right to withdraw from the study at any time without giving any justifications, and do not worry there will be no negative consequences affecting you because of your withdrawal.

Data

The collected data will be visible to me and my supervisor Dr Paul Cairns. This data will be in the form of a spreadsheet, but you are not individually recognisable from the data. The data may be involved in publication. However, in the case of publishing this work, you will not be recognisable in any way.

Participant consent

Your participation in this experiment is entirely voluntary; there will be no remuneration for the time you spend evaluating it. However, your participation is highly appreciated as it will help us understand how related you find different items, and this is important in my research which aims at discovering the factors that affect user performance during menu search.

If you are aware of any medical or other condition that might make it unsafe for you to participate in this study, please do not sign this form or discuss your concerns with the experimenter before doing this experiment.

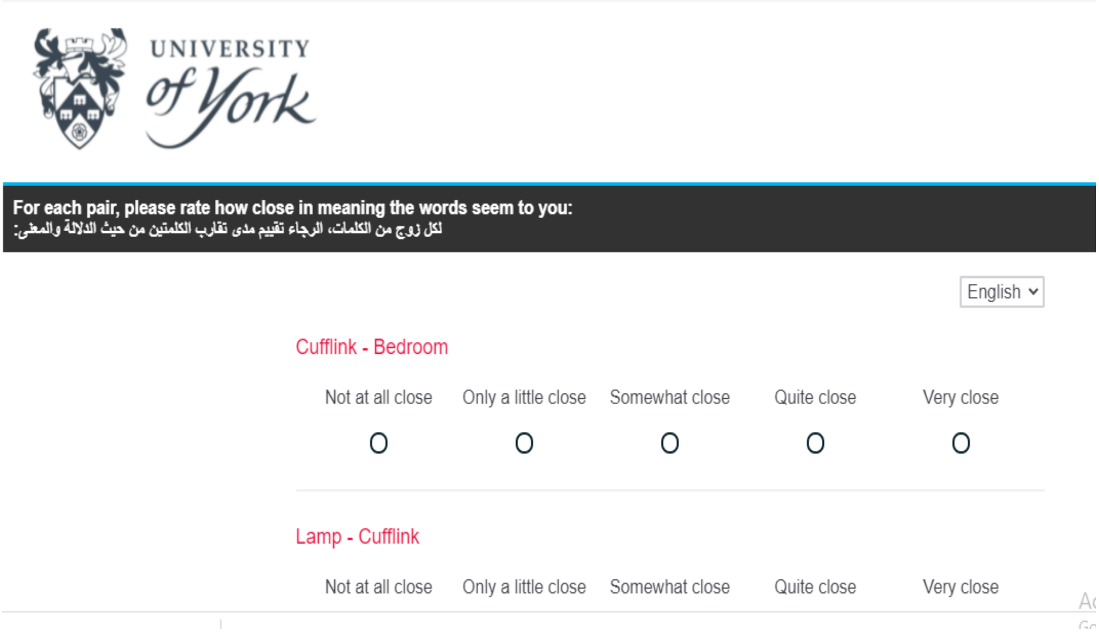
If you are willing to participate, please sign this consent form and proceed with the experiment.

Participant's Signature:

Researcher's contact details:

Researcher: Hend Albassam (haa522@york.ac.uk).
Supervisor: Dr Paul Cairns (p.cairns@yor.ac.uk).

C.2 Study 3 Material (Survey)



UNIVERSITY of York

For each pair, please rate how close in meaning the words seem to you:
لكل زوج من الكلمات، الرجاء تقييم مدى تقارب الكلمتين من حيث الدلالة والمعنى:

English ⌄

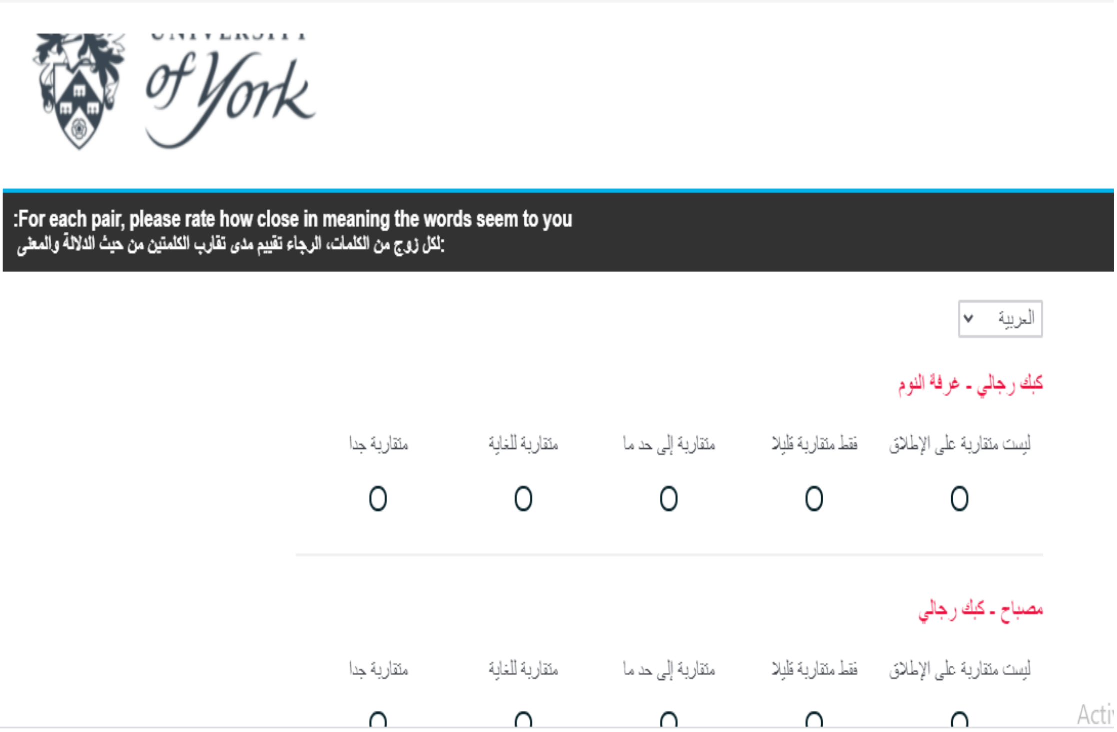
Cufflink - Bedroom

Not at all close	Only a little close	Somewhat close	Quite close	Very close
○	○	○	○	○

Lamp - Cufflink

Not at all close	Only a little close	Somewhat close	Quite close	Very close

English version



UNIVERSITY of York

:For each pair, please rate how close in meaning the words seem to you
لكل زوج من الكلمات، الرجاء تقييم مدى تقارب الكلمتين من حيث الدلالة والمعنى:

العربية ⌄

كبك رجالي - غرفة النوم

متقاربة جدا	متقاربة للغاية	متقاربة إلى حد ما	فقط متقاربة قليلا	ليست متقاربة على الإطلاق
○	○	○	○	○

مصباح - كبك رجالي

متقاربة جدا	متقاربة للغاية	متقاربة إلى حد ما	فقط متقاربة قليلا	ليست متقاربة على الإطلاق
○	○	○	○	○

Arabic version

C.3 Information Sheet and Informed Consent for Study 4

Overview

In this study, you will be asked to group menu items into collections and assign a label for each collection. You will be given a list of menu items, and you should group them according to what you think about how close in meaning these menu items are. There are no right or wrong answers. We are just interested in how similar you find different menu items. In addition, we will gather some demographic data about you. This study will take about 10-15 minutes.

Questions

If you have any questions about the study, please feel free to ask. However, when the study has started, kindly leave your questions until the end.

Withdrawing

You have the right to withdraw from the study at any time without giving any justifications, and do not worry there will be no negative consequences affecting you because of your withdrawal.

Data

The collected data will be visible to me and my supervisor Dr Paul Cairns. This data will be in the form of a spreadsheet, but you not be individually recognisable from the data. The data may be involved in publication. However, in the case of publishing this work, you will not be recognisable in any way.

Participant consent

Your participation in this experiment is entirely voluntary; there will be no remuneration for the time you spend evaluating it. However, your participation is highly appreciated as it will help us understand how related you find different items, and this is important in my research which aims at discovering the factors that affect user performance during menu search.

If you are aware of any medical or other condition that might make it unsafe for you to participate in this study, please do not sign this form or discuss your concerns with the experimenter before doing this experiment.

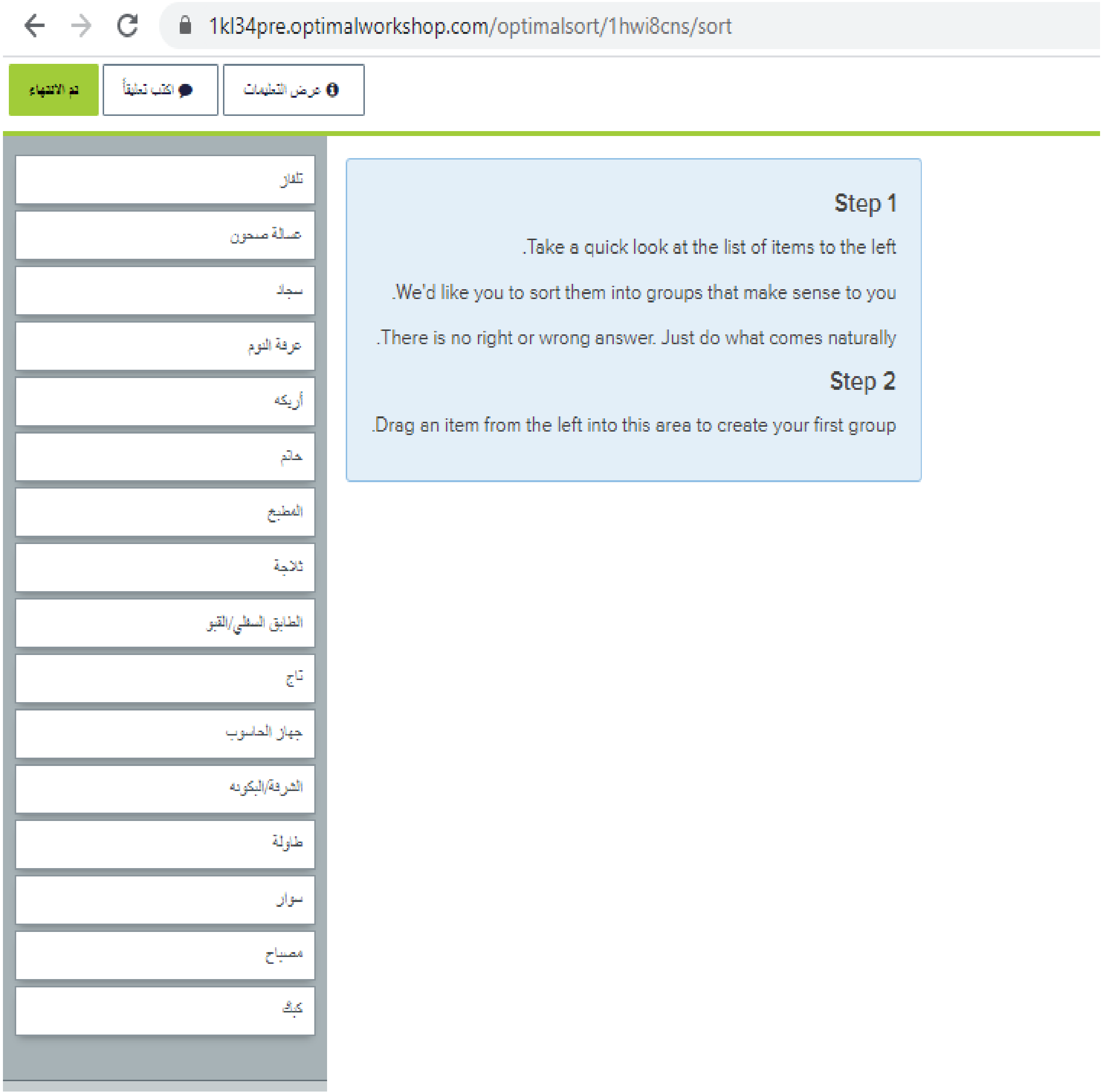
If you are willing to participate, please sign this consent form and proceed with the experiment.

Participant's Signature:

Researcher's contact details:

Researcher: Hend Albassam (haa522@york.ac.uk).
Supervisor: Dr Paul Cairns (p.cairns@yor.ac.uk).

C.4 Study 4 Material (Card Sorting)



Step 1

.Take a quick look at the list of items to the left

.We'd like you to sort them into groups that make sense to you

.There is no right or wrong answer. Just do what comes naturally

Step 2

.Drag an item from the left into this area to create your first group

Arabic version

Appendix D

D.1 The semantic relevance scores of the testing menu samples (the samples that produced outlying performance are highlighted)

Sample#	The semantic relevance scores							
	item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8
1	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
2	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
3	0.166667	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
4	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.5	0.166667
5	0.166667	0.166667	1	0.5	0.166667	0.166667	0.166667	0.166667
6	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
7	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
8	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
9	0.5	0.833333	1	0.5	0.833333	0.166667	0.166667	0.166667
10	0.5	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
11	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
12	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
13	0.833333	0.166667	1	0.833333	0.166667	0.166667	0.166667	0.166667
14	0.5	0.833333	1	0.833333	0.166667	0.5	0.166667	0.166667
15	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
16	0.166667	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.166667
17	0.833333	0.166667	1	0.5	0.166667	0.166667	0.166667	0.166667
18	0.5	0.5	1	0.833333	0.166667	0.166667	0.833333	0.166667
19	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
20	0.5	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
21	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
22	0.5	0.833333	1	0.833333	0.166667	0.166667	0.833333	0.166667
23	0.5	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.166667
24	0.5	0.5	1	0.166667	0.833333	0.166667	0.166667	0.166667
25	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
26	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
27	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
28	0.5	0.5	1	0.5	0.166667	0.166667	0.166667	0.166667
29	0.5	0.833333	1	0.833333	0.833333	0.166667	0.166667	0.166667
30	0.5	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.166667

31	0.833333	0.833333	1	0.5	0.5	0.166667	0.166667	0.166667
32	0.833333	0.5	1	0.833333	0.166667	0.166667	0.833333	0.166667
33	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
34	0.5	0.166667	1	0.833333	0.166667	0.166667	0.166667	0.166667
35	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
36	0.166667	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
37	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
38	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
39	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
40	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
41	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
42	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.833333
43	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
44	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
45	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
46	0.5	0.166667	1	0.833333	0.166667	0.166667	0.166667	0.166667
47	0.5	0.833333	1	0.833333	0.5	0.166667	0.166667	0.166667
48	0.5	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
49	0.5	0.166667	1	0.833333	0.166667	0.166667	0.166667	0.166667
50	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
51	0.833333	0.5	1	0.5	0.166667	0.166667	0.166667	0.166667
52	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
53	0.833333	0.833333	1	0.5	0.166667	0.5	0.166667	0.166667
54	0.5	0.5	1	0.833333	0.166667	0.833333	0.166667	0.166667
55	0.833333	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.166667
56	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
57	0.833333	0.5	1	0.5	0.166667	0.166667	0.166667	0.166667
58	0.833333	0.5	1	0.5	0.166667	0.166667	0.166667	0.5
59	0.833333	0.833333	1	0.5	0.833333	0.166667	0.166667	0.833333
60	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
61	0.833333	0.166667	1	0.833333	0.166667	0.166667	0.166667	0.166667
62	0.833333	0.833333	1	0.833333	0.833333	0.166667	0.166667	0.166667
63	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667

64	0.166667	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
65	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
66	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
67	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
68	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
69	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
70	0.833333	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.166667
71	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
72	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
73	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
74	0.5	0.5	1	0.5	0.166667	0.833333	0.166667	0.166667
75	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
76	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
77	0.833333	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.166667
78	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
79	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
80	0.833333	0.833333	1	0.5	0.166667	0.166667	0.833333	0.166667
81	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
82	0.833333	0.5	1	0.5	0.166667	0.166667	0.166667	0.166667
83	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
84	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
85	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
86	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
87	0.5	0.5	1	0.166667	0.166667	0.166667	0.166667	0.833333
88	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
89	0.833333	0.833333	1	0.833333	0.166667	0.833333	0.166667	0.166667
90	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
91	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
92	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
93	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
94	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
95	0.833333	0.166667	1	0.166667	0.166667	0.833333	0.166667	0.166667
96	0.166667	0.5	1	0.833333	0.833333	0.166667	0.166667	0.166667

97	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
98	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
99	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
100	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.833333

D2. The semantic relevance scores of the testing menu samples (the samples that produced outlying performance are highlighted)

Sample#	The semantic relevance scores							
	item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8
1	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
2	0.166667	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.833333
3	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
4	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
5	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
6	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.166667
7	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
8	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.5
9	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
10	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.833333
11	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.5
12	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.5
13	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
14	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
15	0.166667	0.5	0.166667	0.166667	0.833333	0.833333	1	0.833333
16	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
17	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
18	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.5
19	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
20	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.5
21	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.833333
22	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.5
23	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5

24	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.833333
25	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
26	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
27	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
28	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.5
29	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
30	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
31	0.166667	0.833333	0.166667	0.166667	0.5	0.833333	1	0.833333
32	0.833333	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.166667
33	0.166667	0.833333	0.166667	0.166667	0.833333	0.833333	1	0.833333
34	0.166667	0.833333	0.166667	0.166667	0.833333	0.833333	1	0.833333
35	0.166667	0.166667	0.166667	0.166667	0.166667	0.833333	1	0.5
36	0.166667	0.166667	0.166667	0.5	0.833333	0.833333	1	0.833333
37	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
38	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.5
39	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
40	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
41	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
42	0.166667	0.166667	0.166667	0.5	0.833333	0.833333	1	0.5
43	0.166667	0.166667	0.833333	0.166667	0.5	0.5	1	0.5
44	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.833333
45	0.166667	0.5	0.166667	0.166667	0.833333	0.833333	1	0.833333
46	0.166667	0.166667	0.833333	0.166667	0.833333	0.5	1	0.833333
47	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
48	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
49	0.166667	0.833333	0.166667	0.166667	0.833333	0.833333	1	0.833333
50	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
51	0.166667	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.5
52	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
53	0.166667	0.833333	0.166667	0.166667	0.5	0.833333	1	0.5
54	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
55	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
56	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333

57	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
58	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
59	0.5	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
60	0.166667	0.166667	0.166667	0.166667	0.166667	0.833333	1	0.5
61	0.166667	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.833333
62	0.5	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
63	0.166667	0.166667	0.166667	0.166667	0.166667	0.166667	1	0.5
64	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
65	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
66	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.166667
67	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.5
68	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
69	0.166667	0.166667	0.166667	0.833333	0.833333	0.833333	1	0.833333
70	0.833333	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
71	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
72	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
73	0.166667	0.166667	0.166667	0.5	0.833333	0.833333	1	0.833333
74	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.5
75	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
76	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
77	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
78	0.166667	0.166667	0.166667	0.833333	0.833333	0.833333	1	0.833333
79	0.166667	0.166667	0.833333	0.833333	0.5	0.5	1	0.833333
80	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
81	0.166667	0.166667	0.833333	0.166667	0.833333	0.833333	1	0.5
82	0.166667	0.166667	0.166667	0.5	0.833333	0.833333	1	0.833333
83	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
84	0.166667	0.166667	0.5	0.166667	0.833333	0.5	1	0.833333
85	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.166667
86	0.833333	0.166667	0.166667	0.166667	0.833333	0.5	1	0.5
87	0.833333	0.166667	0.166667	0.166667	0.833333	0.5	1	0.5
88	0.166667	0.166667	0.166667	0.833333	0.833333	0.166667	1	0.5
89	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5

90	0.166667	0.166667	0.166667	0.166667	0.166667	0.5	1	0.5
91	0.166667	0.833333	0.166667	0.166667	0.5	0.833333	1	0.833333
92	0.166667	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.5
93	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
94	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.5
95	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
96	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
97	0.166667	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.833333
98	0.5	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.5
99	0.166667	0.166667	0.166667	0.833333	0.833333	0.833333	1	0.5
100	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5

D3. The semantic relevance scores of the testing menu samples (the samples that produced outlying performance are highlighted)

Sample#	The semantic relevance scores							
	item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8
1	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
2	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
3	0.5	0.833333	1	0.5	0.166667	0.833333	0.166667	0.166667
4	0.833333	0.5	1	0.5	0.166667	0.166667	0.166667	0.166667
5	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
6	0.5	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
7	0.833333	0.5	1	0.5	0.166667	0.166667	0.166667	0.166667
8	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
9	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
10	0.833333	0.5	1	0.5	0.166667	0.166667	0.166667	0.166667
11	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
12	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
13	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
14	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
15	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
16	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667

17	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
18	0.833333	0.5	1	0.833333	0.5	0.166667	0.166667	0.166667
19	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
20	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
21	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
22	0.833333	0.833333	1	0.833333	0.166667	0.833333	0.166667	0.166667
23	0.833333	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.5
24	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.833333	0.166667
25	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
26	0.833333	0.833333	1	0.833333	0.833333	0.166667	0.166667	0.166667
27	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
28	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
29	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
30	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
31	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
32	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
33	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
34	0.833333	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.166667
35	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
36	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
37	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
38	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
39	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
40	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
41	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
42	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
43	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
44	0.833333	0.5	1	0.5	0.166667	0.166667	0.166667	0.166667
45	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
46	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
47	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
48	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
49	0.5	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667

50	0.5	0.833333	1	0.166667	0.166667	0.166667	0.833333	0.166667
51	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
52	0.5	0.166667	1	0.833333	0.166667	0.166667	0.166667	0.166667
53	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
54	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
55	0.5	0.166667	1	0.5	0.166667	0.166667	0.166667	0.166667
56	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
57	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
58	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
59	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.833333	0.166667
60	0.166667	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
61	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
62	0.833333	0.833333	1	0.166667	0.166667	0.166667	0.166667	0.166667
63	0.5	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
64	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
65	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
66	0.5	0.5	1	0.833333	0.166667	0.166667	0.833333	0.166667
67	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
68	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
69	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
70	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
71	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
72	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
73	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
74	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
75	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
76	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
77	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
78	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
79	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
80	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
81	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
82	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667

83	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
84	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
85	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
86	0.5	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
87	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
88	0.833333	0.833333	1	0.5	0.166667	0.166667	0.166667	0.166667
89	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
90	0.166667	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
91	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
92	0.166667	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
93	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
94	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
95	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
96	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
97	0.833333	0.5	1	0.833333	0.166667	0.166667	0.166667	0.166667
98	0.833333	0.166667	1	0.5	0.166667	0.166667	0.166667	0.166667
99	0.833333	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667
100	0.5	0.833333	1	0.833333	0.166667	0.166667	0.166667	0.166667

D4. The semantic relevance scores of the testing menu samples (the samples that produced outlying performance are highlighted)

Sample#	The semantic relevance scores							
	item 1	item 2	item 3	item 4	item 5	item 6	item 7	item 8
1	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
2	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
3	0.166667	0.166667	0.166667	0.833333	0.833333	0.833333	1	0.833333
4	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
5	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
6	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.5
7	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333

8	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
9	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
10	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
11	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
12	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
13	0.166667	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.833333
14	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
15	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
16	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
17	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
18	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.5
19	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.5
20	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
21	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
22	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
23	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
24	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
25	0.166667	0.166667	0.166667	0.166667	0.166667	0.833333	1	0.5
26	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
27	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
28	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
29	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
30	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
31	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
32	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
33	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
34	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
35	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
36	0.166667	0.166667	0.833333	0.166667	0.833333	0.5	1	0.833333
37	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
38	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
39	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
40	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333

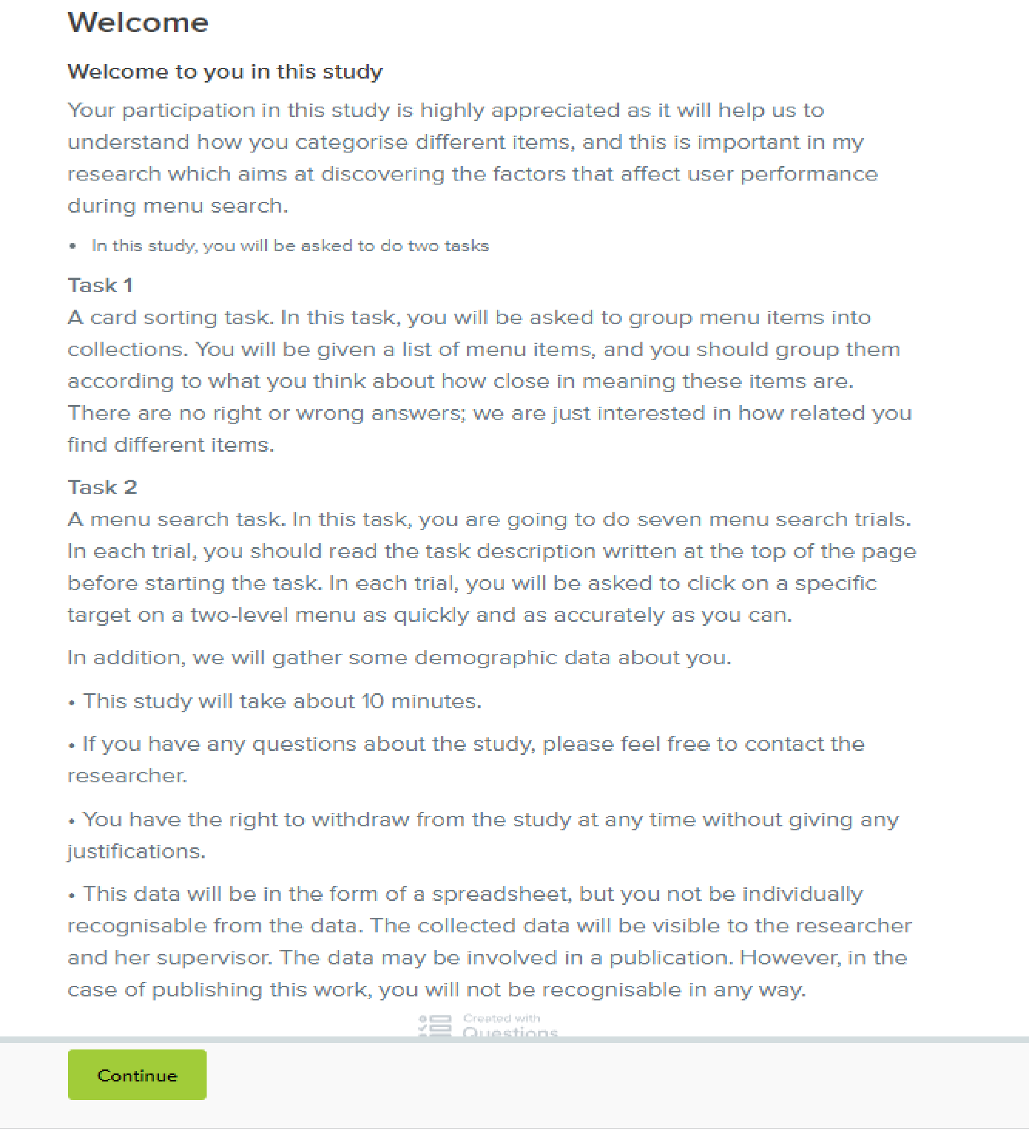
41	0.166667	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.833333
42	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
43	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
44	0.166667	0.166667	0.166667	0.833333	0.833333	0.833333	1	0.5
45	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
46	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
47	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
48	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
49	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
50	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
51	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
52	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
53	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
54	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
55	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
56	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
57	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
58	0.166667	0.166667	0.166667	0.5	0.833333	0.166667	1	0.5
59	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
60	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
61	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
62	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
63	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
64	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
65	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
66	0.166667	0.166667	0.166667	0.166667	0.5	0.166667	1	0.833333
67	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
68	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
69	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
70	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
71	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
72	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
73	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5

74	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
75	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.5
76	0.166667	0.166667	0.166667	0.5	0.166667	0.833333	1	0.833333
77	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
78	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.5
79	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
80	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
81	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
82	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
83	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.5
84	0.166667	0.166667	0.166667	0.166667	0.833333	0.166667	1	0.833333
85	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
86	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
87	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
88	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
89	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.5
90	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
91	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
92	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
93	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.5
94	0.166667	0.166667	0.166667	0.166667	0.5	0.5	1	0.833333
95	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
96	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5
97	0.166667	0.166667	0.166667	0.166667	0.5	0.833333	1	0.833333
98	0.166667	0.166667	0.166667	0.166667	0.833333	0.5	1	0.833333
99	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.833333
100	0.166667	0.166667	0.166667	0.166667	0.833333	0.833333	1	0.5

Appendix E

Chapter 7

E.1 Information Sheet and Informed Consent for Study 6



Welcome

Welcome to you in this study

Your participation in this study is highly appreciated as it will help us to understand how you categorise different items, and this is important in my research which aims at discovering the factors that affect user performance during menu search.

- In this study, you will be asked to do two tasks

Task 1
A card sorting task. In this task, you will be asked to group menu items into collections. You will be given a list of menu items, and you should group them according to what you think about how close in meaning these items are. There are no right or wrong answers; we are just interested in how related you find different items.

Task 2
A menu search task. In this task, you are going to do seven menu search trials. In each trial, you should read the task description written at the top of the page before starting the task. In each trial, you will be asked to click on a specific target on a two-level menu as quickly and as accurately as you can.

In addition, we will gather some demographic data about you.

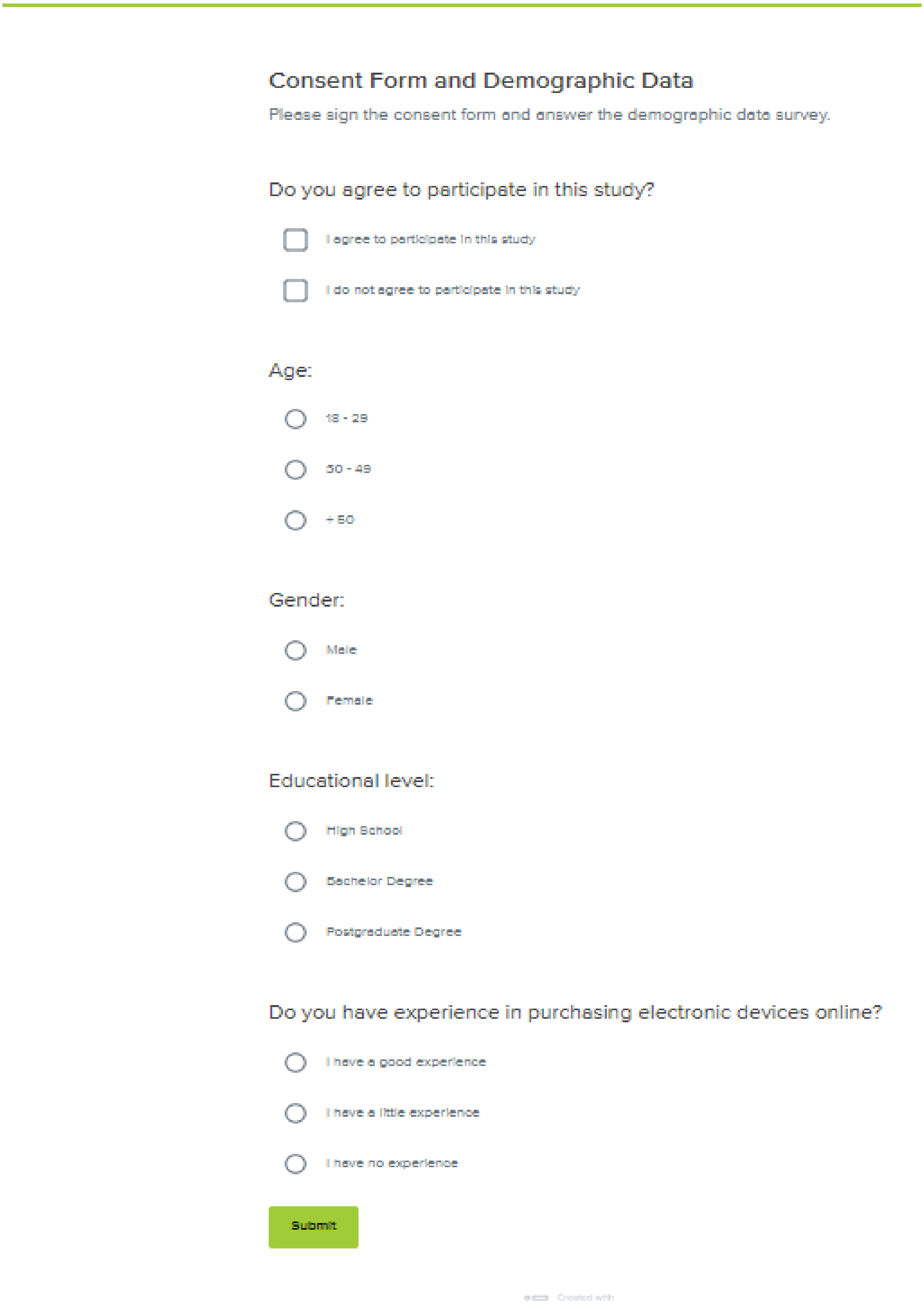
• This study will take about 10 minutes.

• If you have any questions about the study, please feel free to contact the researcher.

• You have the right to withdraw from the study at any time without giving any justifications.

• This data will be in the form of a spreadsheet, but you not be individually recognisable from the data. The collected data will be visible to the researcher and her supervisor. The data may be involved in a publication. However, in the case of publishing this work, you will not be recognisable in any way.

Information Sheet



Consent Form and Demographic Data

Please sign the consent form and answer the demographic data survey.

Do you agree to participate in this study?

□ I agree to participate in this study

□ I do not agree to participate in this study

Age:

○ 18 - 29

○ 30 - 49

○ + 50

Gender:

○ Male

○ Female

Educational level:

○ High School

○ Bachelor Degree

○ Postgraduate Degree

Do you have experience in purchasing electronic devices online?

○ I have a good experience

○ I have a little experience

○ I have no experience

Consent Form and Demographic Data

Appendix F

Chapter 8

F1. Information Sheet and Informed Consent for Study 7

Overview

In this study, we are interested in understanding user behaviour during menu search tasks. Therefore, you will be asked to carry out a menu search task in silence and without any assistance from the researcher. Your performance in the task will be recorded using screen capture software. After completing the task, you will sit with the researcher, and you will be shown the recording of your performance in the menu search task and asked to comment on the process in retrospect, and your voice will be recorded. In addition, we will gather some demographic data about you. This study will take about 15-20 minutes.

Questions

If you have any questions about the study, please feel free to ask. However, when the study has started, kindly leave your questions until the end.

Withdrawing

You have the right to withdraw from the study at any time without giving any justifications, and do not worry there will be no negative consequences affecting you because of your withdrawal. Recordings and notes taken will be destroyed as you require.

Data

The collected data will be visible to me and my supervisor Dr Paul Cairns. The collected audio and video data will be deleted after analysing them. The data may be involved in publication. However, in the case of publishing this work, you will not be recognisable in any way.

Participant consent

Your participation in this study is entirely voluntary; there will be no remuneration for the time you spend evaluating it. However, your participation is highly appreciated as it will help us to understand the user behaviour during menu search tasks. During the study, it will be necessary for me to record a number of things using screen capture software and a voice recorder. However, this recorded data will be stored securely on the university approved cloud in accordance to the University of York data protection policy.

If you are aware of any medical or other condition that might make it unsafe for you to participate in this study, please do not participate in this study. There will be no penalty or negative consequences for not participating in this study.

If you are willing to participate, please sign this consent form and proceed with the experiment.

Participant's Signature:

Researcher's contact details:

Researcher: Hend Albassam (haa522@york.ac.uk).

Supervisor: Dr Paul Cairns (p.cairns@yor.ac.uk).

11. References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods, 16*(2), 270–301. https://doi.org/10.1177/1094428112470848.

Albert, W., Tullis, T., & Tedesco, D. (2010). *Beyond the Usability Lab: Conducting Large scale Online User Experience Studies* (1st ed.). Morgan Kaufmann.

Albert, W., & Tullis, T. (2013). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics (Interactive Technologies)* (2nd ed.). Morgan Kaufmann.

Alnashri, A., Alhadreti, O., & Mayhew, P. J. (2016). The Influence of Participant Personality in Usability Tests. *International Journal of Human-Computer Interaction (IJHCI)*, *7*(1).

Auskerin, A. (2012). *Investigating Really Poor User Performance* [Unpublished BEng dissertation]. University of York, York, UK.

Bailly, G., Lecolinet, E., & Nigay, L. (2008). Flower menus: A new type of Marking menu with large menu breadth, within groups and efficient expert mode memorization. *Proceedings of the Workshop on Advanced Visual Interfaces AVI*. 15-22. https://doi.org/10.1145/1385569.1385575.

Bailly, G., Oulasvirta, A., Brumby, D. P., & Howes, A. (2014). Model of visual search and selection time in linear menus. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3865-3874). ACM. https://doi.org/10.1145/2556288.2557093.

Bailly, G., Lecolinet, E., & Nigay, L. (2016). Visual Menu Techniques. *ACM Computing Surveys*, *49*(4), 1–41. https://doi.org/10.1145/3002171.

BALOH, R. W., SILLS, A. W., KUMLEY, W. E., & HONRUBIA, V. (1975). Quantitative measurement of saccade amplitude, duration, and velocity. *Neurology*, *25*(11), 1065. https://doi.org/10.1212/wnl.25.11.1065.

Barnum, C. M. (2011). *Usability Testing Essentials: Ready, Set . . .Test!: Ready, Set. . .Test!*. Morgan Kaufmann.

Boren, T., & Ramey, J. (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication*, *43*(3), 261–278. https://doi.org/10.1109/47.867942.

Brasoveanu A., Dotlačil J. (2020). The ACT-R Cognitive Architecture and Its pyactr Implementation. In: *Computational Cognitive Modeling and Linguistic Theory* (pp. 7-37). Springer. https://doi.org/10.1007/978-3-030-31846-8_2.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa.

Brumby, D. P., & Howes, A. (2003). Interdependence and past experience in menu choice assessment. *Proceedings of the Annual Meeting of the Cognitive Science Society, 25(25), 1320.* https://escholarship.org/content/qt4nm1z48r/qt4nm1z48r.pdf.

Brumby, D. P., Cox, A. L., Chung, J., & Fernandes, B (2014). How does knowing what you are looking for change visual search behavior? *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3895-3898). ACM. https://doi.org/10.1145/2556288.2557064.

Brumby, D. P., & Zhuang, S. (2015). Visual grouping in menu interfaces. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4203-4206). ACM. https://doi.org/10.1145/2702123.2702177.

Burleigh, T. (2019, August 4). *What is fair payment on MTurk?* https://tylerburleigh.com/blog/what-is-fair-payment-on-mturk/

Burnett, G. E., & Ditsikas, D. (2006). Personality as a criterion for selecting usability testing participants. In *Proc. int. conf. on information and* *communications technologies* (pp. 599-604). IEEE. https://doi.org/10.1109/ITICT.2006.358235.

Byrene, M. D. (2001). ACT-R/PM and menu selection: applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies*, *55*(1), 41–84. https://doi.org/10.1006/ijhc.2001.0469.

Cairns, P. (2007). HCI... not as it should be: inferential statistics in HCI research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it* (pp. 195-201). British Computer Society. https://doi.org/10.14236/ewic/HCI2007.20.

Cairns, P., & Cox, A. L. (2008). *Research Methods for Human-Computer Interaction* (Illustrated ed.). Cambridge University Press.

Cairns, P. (2019). *Doing Better Statistics in Human-Computer Interaction*. Cambridge University Press.

Card, S. K. (1982). User perceptual mechanisms in the search of computer command menus. *Proceedings of the 1982 conference on Human factors in computing systems*.190-196. ACM. https://doi.org/10.1145/800049.801779.

Charles, W. G. (2000). Contextual correlates of meaning. *Applied Psycholinguistics*, *21*(4), 505–524. https://doi.org/10.1017/s0142716400004057.

Chen, X., Bailly, G., Brumby, D. P., Oulasvirta, A., & Howes, A. (2015). The emergence of interactive behaviour: A model of rational menu search. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4217-4226. ACM. https://doi.org/10.1145/2702123.2702483.

Chen, X. (2015). *An optimal control approach to testing theories of human information processing constraints* [PhD thesis, University of Birmingham]. ETHOS. https://ethos.bl.uk/OrderDetails.do?did=1&uin=uk.bl.ethos.649294.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2016). Amazon Mechanical Turk in Organizational Psychology: An Evaluation and Practical Recommendations. *Journal of Business and Psychology*, *32*(4), 347–361. https://doi.org/10.1007/s10869-016-9458-5.

Chua, S. L., Chen, D. T., & Wong, A. F. (1999). Computer anxiety and its correlates: a meta-analysis. *Computers in Human Behavior*, *15*(5), 609–623. https://doi.org/10.1016/s0747-5632(99)00039-4.

Cockburn, A., Gutwin, C., & Greenberg, S. (2007). A predictive model of menu performance. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 627-636. ACM. https://doi.org/10.1145/1240624.1240723.

Cotton, D., & Gresty, K. (2006). Reflecting on the think-aloud method for evaluating e-learning. *British Journal of Educational Technology, 37*(1), 45–54. https://doi.org/10.1111/j.1467-8535.2005.00521.x.

Cox, A. L., & Young, R. M. (2004). A Rational Model of the Effect of Information Scent on the Exploration of Menus. *Proceedings of the Sixth International Conference on Cognitive Modelling.* 82-87. Citeseerx. https://citeseerx.ist.psu.edu/viewdoc/citations?doi=10.1.1.107.8819 .

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410.

Dachselt, R., & Hübner, A. (2007). Three-dimensional menus: A survey and taxonomy. *Computers & Graphics*, *31*(1), 53–65. https://doi.org/10.1016/j.cag.2006.09.006.

Deibel, K., Anderson, R., & Anderson, R. (2005). Using edit distance to analyze card sorts. *Expert Systems*, *22*(3), 129–138. https://doi.org/10.1111/j.1468-0394.2005.00304.x

Dillon, A., & Watson, C. (1996). User analysis in HCI — the historical lessons from individual differences research. *International Journal of Human-*

Computer Studies, *45*(6), 619–637. https://doi.org/10.1006/ijhc.1996.0071.

Dumas, J. S., & Redish, J. C. (1993). *A Practical Guide to Usability Testing.* Ablex Publishing Corporation.

Dwyer, T. J. (2003). *An Assessment of Paired Similarities and Card Sorting.* [Master thesis, University of South Florida]. https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=2358&context=etd.

Dyck, J. L., & Smither, J. A. A. (1994). Age Differences in Computer Anxiety: The Role of Computer Experience, Gender and Education. *Journal of Educational Computing Research*, *10*(3), 239–248. https://doi.org/10.2190/e79u-vcrc-el4e-hryv.

Egan, D. E. (1988). Individual differences in human-computer interaction. In *Handbook of human-computer interaction*. (pp. 543-568). North-Holland. https://doi.org/10.1016/B978-0-444-70536-5.50029-4.

Ehrler, F., Weinhold, T., Joe, J., Lovis, C., & Blondon, K. (2018). A Mobile App (BEDSide Mobility) to Support Nurses' Tasks at the Patient's Bedside: Usability Study. *JMIR mHealth and uHealth*, *6*(3), e57. https://doi.org/10.2196/mhealth.9079.

Ericsson, A. K., & Simon, H. A. (1993). *Protocol Analysis - Rev'd Edition: Verbal Reports as Data* (revised edition). A Bradford Book.

Faiks, A., & Hyland, N. (2000). Gaining User Insight: A Case Study Illustrating the Card Sort Technique. *College & Research Libraries*, *61*(4), 349–357. https://doi.org/10.5860/crl.61.4.349.

Fergus, T. A., & Rowatt, W. C. (2014). Intolerance of uncertainty and personality: Experiential permeability is associated with difficulties tolerating uncertainty. *Personality and Individual Differences, 58,* 128–131. https://doi.org/10.1016/j.paid.2013.10.017.

Følstad, A., Law, E. L. C., & Hornbæk, K. (2012). Outliers in usability testing: How to treat usability problems found for only one test participant?

Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design. 257-260. ACM. https://doi.org/10.1145/2399016.2399056.

Fontana, A., & Frey, J. H. (2005). The Interview: From Neutral Stance to Political Involvement. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (pp. 695–727). Sage Publications Ltd.

Galarnyk, M. (2018, Sep 12). *Understanding Boxplots - Towards Data Science*. Medium. https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51.

Glasser, A. (2019, May). Automatic speech recognition services: deaf and hard-of-hearing usability. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-6).

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye-tracking. *Proceedings of the SIGCHI Conference on Human Factors in computing systems.* 1253-1262. ACM. https://doi.org/10.1145/1124772.1124961.

Halverson, T., & Hornof, A. J. (2008). The effects of semantic grouping on visual search. *CHI'08 Extended Abstracts on Human Factors in Computing Systems.* 3471-3476. ACM. https://doi.org/10.1145/1358628.1358876.

Helander, M., Landauer, T. K., Prabhu, P. V. (1997). *Handbook of human-computer interaction* (2nd ed.). Elsevier.

Hillmer, B. (2020). *Randomize Questions*. Alchemer. https://help.alchemer.com/help/randomize-questions.

Hornbæk, K., & Law, E. L. C. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 617-626. ACM. https://doi.org/10.1145/1240624.1240722.

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology*, *29*(1), 97–111. https://doi.org/10.1080/01449290801939400.

Hornof, A. J., & Kieras, D. E. (1999). Cognitive modelling demonstrates how people use anticipated location knowledge of menu items. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 410-417. ACM. https://doi.org/10.1145/302979.303120.

Howell, D. C. (2016). *Fundamental Statistics for the Behavioral Sciences* (9th ed.). Cengage Learning.

Iftikhar, A., Bond, R. R., McGilligan, V., Leslie, S. J., Rjoob, K., Knoery, C., Quigg, C., Campbell, R., Boyd, K., McShane, A., & Peace, A. (2021). Comparing Single-Page, Multipage, and Conversational Digital Forms in Health Care: Usability Study. *JMIR Human Factors*, *8*(2), e25787. https://doi.org/10.2196/25787.

ISO. (2018). *ISO 9241-11:2018(en) Ergonomics of human-system interaction* — Part 11: Usability: Definitions and concepts. ISO. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

Jacoby, R. H., & Ellis, S. R. (1992). Using virtual menus in a virtual environment. *Proceedings of SPIE - The International Society for Optical Engineering*. https://doi.org/10.1117/12.59654.

Jakobsen, M. R., & Hornæk, K. (2007). Transient visualizations. *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces* (pp. 69-76). https://doi.org/10.1145/1324892.1324905.

Kassambara, A. (2017). *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (Multivariate Analysis) (Volume 1)* (1st ed.). CreateSpace Independent Publishing Platform.

Kieras, D. E., & Hornof, A. J. (2014). Towards accurate and practical predictive models of active-vision-based visual search. *Proceedings of the SIGCHI* *conference on human factors in computing systems*. 3875-3884. ACM. https://doi.org/10.1145/2556288.2557324.

Kjeldskov, J., Skov, M. B., & Stage, J. (2004). Instant data analysis: conducting usability evaluations in a day. *Proceedings of the third Nordic conference on Human-computer interaction.* 233-240. ACM. https://doi.org/10.1145/1028014.1028050.

Kroemer, K. K. H. B. (2000). *Ergonomics: How to Design for Ease and Efficiency:2nd (Second) edition*. Prentice Hall.

Kuurstra, J. (2015). Individual differences in Human-Computer Interaction: A review of empirical studies. [Master thesis, University of Twente]. http://essay.utwente.nl/68638/1/Kuurstra%2C%20J.%20-%20s1099221%20%28verslag%29.pdf.

Kvale, S. (2007). *Doing Interviews*. SAGE Publications. https://www.doi.org/10.4135/9781849208963.

Lantz, E., Keeley, J. W., Roberts, M. C., Medina-Mora, M. E., Sharan, P., & Reed, G. M. (2019). Card Sorting Data Collection Methodology: How Many Participants Is Most Efficient? *Journal of Classification, 36*(3), 649–658. https://doi.org/10.1007/s00357-018-9292-8.

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods* (1st ed.). SAGE Publications, Inc. https://doi.org/10.4135/9781412963947.

Law, E. L. C., & Hvannberg, E. T. (2004). Analysis of combinatorial user effect in international usability tests. *In Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 9-16). ACM. https://doi.org/10.1145/985692.985694.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research Methods in Human-Computer Interaction* (2nd ed.). Morgan Kaufmann.

Lee, E., & Macgregor, J. (1985). Minimizing User Search Time in Menu Retrieval Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *27*(2), 157–162. https://doi.org/10.1177/001872088502700203.

Lee, E. S., & Raymond, D. R. (1993). Menu-driven systems. *Encyclopedia of Microcomputers*, *11*, 101-127.

Lewis, J. R. (2006). Sample sizes for usability tests. *Interactions*, *13*(6), 29–33. https://doi.org/10.1145/1167948.1167973.

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. *International Review of Social Psychology*, *32*(1). https://doi.org/10.5334/irsp.289

MacKenzie, S. I. (2013). *Human-Computer Interaction: An Empirical Research Perspective* (1st ed.). Morgan Kaufmann.

Maguire, M., & Delahunt, B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland Journal of Higher Education*, *9*(3). https://ojs.aishe.org/index.php/aishe-j/article/view/335.

Matthews, G., Davies, D. R., Westerman, S. J., & Stammers, R. B. (2000). *Human performance. Cognition, stress, and individual differences.* Psychology Press.

Mayer, C., Hanenberg, S., Robbes, R., Tanter, R., & Stefik, A. (2012). An empirical study of the influence of static type systems on the usability of undocumented software. *ACM SIGPLAN Notices*, *47*(10), 683–702. https://doi.org/10.1145/2398857.2384666.

McCartan, K., & Robson, C. (2016). *Real World Research* (4th ed.). Wiley.

McClelland, G. H. (2000). Nasty data: Unruly, ill-mannered observations can ruin your analysis. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 393–411). Cambridge University Press.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*(1), 1–28. https://doi.org/10.1080/01690969108406936.

Mtonga, T., Abaye, M., Rosko, S. C., & Douglas, G. P. (2018). A comparative usability study of two touchscreen clinical workstations for use in low resource settings. *Journal of Health Informatics in Africa, 5*(2).

Nakhimovsky, Y., Schusteritsch, R., & Rodden, K. (2006). Scaling the card sort method to over 500 items: restructuring the Google AdWords Help Center. *CHI'06 Extended Abstracts on Human Factors in Computing Systems* (pp. 183-188). ACM. https://doi.org/10.1145/1125451.1125491.

Nielsen, J. (1993). *Usability Engineering* (1st ed.). Morgan Kaufmann.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *In Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206-213). ACM. https://doi.org/10.1145/169059.169166.

Nielsen, J. (2000). *Why you only need to test with 5 users*. nngroup. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/.

Nielsen, J (2006). *Corporate UX Maturity: Stages 1-4*. nngrroup. https://www.nngroup.com/articles/ux-maturity-stages-1-4/.

Norman, D. (2013). *The Design of Everyday Things: Revised and Expanded Edition* (Revised ed.). Basic Books.

Norman, K. (1992). The psychology of menu selection: Designing cognitive control of the Human/Computer Interface. *Displays, 13*(4), 206. https://doi.org/10.1016/0141-9382(92)90066-z.

Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010). Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2381-2390). ACM. https://doi.org/10.1145/1753326.1753685.

O'Shea, J., Bandar, Z., Crockett, K., & McLean, D. (2010). Benchmarking short text semantic similarity. *International Journal of Intelligent Information* *and Database Systems*, *4*(2), 103. https://doi.org/10.1504/ijiids.2010.032437.

Payne, S. J., & Howes, A. (2013). *Adaptive Interaction: A Utility Maximization Approach to Understanding Human Interaction with Technology (Synthesis Lectures on Human-Centered Informatics)* (Illustrated ed.). Morgan & Claypool Publishers.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643–675. https://doi.org/10.1037/0033-295x.106.4.643.

Preece, J., Sharp, H., & Rogers, Y. (2015). *Interaction Design: Beyond Human-Computer Interaction* (4th ed.). Wiley.

Raskin, J. (2000). *The Humane Interface: New Directions for Designing Interactive Systems*. Addison-Wesley Professional.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, *11*, 95–130. https://doi.org/10.1613/jair.514.

Robertson, I., & Kortum, P. (2020). Validity of Three Discount Methods for Measuring Perceived Usability. *Journal of Usability Studies*, *16*(1).

Rubin, J., & Chisnell, D. (2008). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. John Wiley & Sons.

Salah, D., Paige, R., & Cairns, P. (2014). Integrating agile development processes and user centred design-a place for usability maturity models? *International Conference on Human-Centred Software Engineering* (pp. 108-125). Springer. https://doi.org/10.1007/978-3-662-44811-3_7.

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1609-1618). ACM. https://doi.org/10.1145/1518701.1518947.

Sauro, J., & Lewis, J. R. (2010). Average task times in usability tests: what to report? *Proceedings of the SIGCHI conference on human factors in* *computing systems* (pp. 2347-2350). ACM. https://doi.org/10.1145/1753326.1753679.

Sauro, J. (2011). *10 Things To Know About Task Times*. Measuring U. https://measuringu.com/task-times/.

Schiller, J., & Cairns, P. (2008). There's always one!: modelling outlying user performance. *CHI'08 Extended Abstracts on Human Factors in Computing Systems* (pp. 3513-3518). ACM. https://doi.org/10.1145/1358628.1358883.

Schmettow, M., & Sommer, J. (2016). Linking card sorting to browsing performance – are congruent municipal websites more efficient to use? *Behaviour & Information Technology, 35*(6), 452–470. https://doi.org/10.1080/0144929x.2016.1157207.

Schmidt, T., Wittmann, V., & Wolff, C. (2019). The Influence of Participants' Personality on Quantitative and Qualitative Metrics in Usability Testing. *Proceedings of Mensch und Computer 2019* (pp. 115-126). ACM. https://doi.org/10.1145/3340764.3340787.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99. https://doi.org/10.2307/1884852

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets* [Doctoral dissertation, University of Pittsburgh]. http://d-scholarship.pitt.edu/7948/.

Shneiderman, B. (2000b). Universal usability. *Communications of the ACM, 43*(5), 84–91. https://doi.org/10.1145/332833.332843

Sonderegger, A., Schmutz, S., & Sauer, J. (2016). The influence of age in usability testing. *Applied Ergonomics, 52,* 291–300. https://doi.org/10.1016/j.apergo.2015.06.012.

Spencer, D. (2009). *Card Sorting: Designing Usable Categories* (1st ed.). Rosenfeld Media.

Stein, R., & Swan, A. B. (2019). Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass*, *13*(2), e12434. https://doi.org/10.1111/spc3.12434.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning* (Vol. 135). Cambridge: MIT press.

Tullis, T. S. (1985). Designing a menu-based interface to an operating system. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 79-84). ACM. https://doi.org/10.1145/317456.317471.

Wilcox, R. R. (2014). *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy* (2nd ed. 2010 ed.). Springer.

Wood, J., Wood, L. (2008). Card sorting: current practices and beyond. *Journal of Usability Studies*, 4(1), 1-6. https://doi.org/10.5555/2835577.2835578.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. *Proceedings of IHM-HCI 2001 conference* (pp. 105-108). Citeseerx. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.85.7896.

Yin, Z. (2018). *Investigating the outlying performance in menu search* [Unpublished master's thesis]. University of York, York, UK.

Yoon, C., Feinberg, F., Hu, P., Gutchess, A. H., Hedden, T., Chen, H. Y. M., Jing, Q., Cui, Y., & Park, D. C. (2004). Category Norms as a Function of Culture and Age: Comparisons of Item Responses to 105 Categories by American and Chinese Adults. *Psychology and Aging*, *19*(3), 379–393. https://doi.org/10.1037/0882-7974.19.3.379.

Young, R. M. (1998). Rational analysis of exploratory choice. *Rational models of cognition*, 469-500.

van den Haak, M., de Jong, M., & Jan Schellens, P. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*(5), 339–351. https://doi.org/10.1080/0044929031000.

van der Kloot, W. A., & van Herk, H. (1991). Multidimensional Scaling of Sorting Data: A Comparison of Three Procedures. *Multivariate Behavioral Research*, *26*(4), 563–581. https://doi.org/10.1207/s15327906mbr2604_1.