City University London

MSc in Human-Centred Systems

Project Report

2015

Gender differences in using recommender systems

Paul Galbraith

Supervised by: Dr. Simone Stumpf

25th September 2015

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: 

Abstract

The focus of this research is in the area of gender differences in the use of intelligent interactive systems, for which recommender systems are included. Such a study is important for if there are gender differences found for these low-cost systems, then it could impact a high number of the population. The research includes a review of literature which has identified gender differences, and their causes. An experiment took place involving males and females interacting with a TV recommender system. One significant result was found, along with a small number of marginal results. The main conclusion being, that even with a low-cost system, there were differences. This dissertation recommends the need for more research, involving higher numbers of participants.

Keywords: gender, recommender systems, intelligent interactive systems

Contents

Chapter 1: Introduction and objectives

1.1 Background

Over the past 20 years, recommender systems (also known as recommender engines) have grown in popularity to the point that they are now ubiquitous. This is evident to anyone who spends much time online—sooner or later they will end up visiting a website (or downloading an app) that uses some type of recommender system. Consider some of the most popular websites (Wikipedia, 2015), from companies such as Amazon, Apple, Facebook, Google, Netflix, Twitter, Wikipedia and YouTube—they all use recommender systems. These types of systems provide recommendations for all manner of things, such as what to watch or read, with whom to be friends, who to follow, what to buy. With so many different recommender systems in use now, it is hard to imagine many decisions for which one of these systems could not supply useful recommendations.

However, it has been found with other intelligent interactive systems, for which recommendation systems are a subclass, that the way in which males and females interact with these systems can vary significantly. Theories exist for why there are these gender differences, explained by so-called psychological constructs, with risk aversion being one such example. Risk aversion can be described as a behavioural reaction to uncertainty, with females found to be more risk-averse than males (Croson and Gneezy, 2009). Other examples of where gender differences have been found, include: information processing theory (Meyers-Levy, 1989), and self-efficacy (Beckwith *et al.*, 2005). It could therefore be expected that these known psychological constructs are also influential in how recommender systems are used by males and females, thus resulting in gender differences.

Knowing whether gender differences exist in the use of a particular type of system, and what those differences are, is important. As it could be the case that the design of the system is more beneficial to males than females, or vice versa. Therefore, once this information is known, changes can be made to the system—with the intention of making it equally effective for both genders (Grigoreanu *et al.*, 2008).

1.2 Research focus

An aim for research is to add to the knowledge that already exists in terms of the types of systems and what gender differences have been reported in the use of them. A number of studies have looked at gender in relation to the use of intelligent interactive systems, machine-learning, and other types of computer software—of which, pertinent titles will be discussed in the literature review. However, studies that look for gender differences in the use of recommender systems are harder to find, although there are some interesting studies involving just the use of recommender systems, and related systems, that will provide useful insights.

This study will investigate whether gender differences exist in the use of recommender systems, and if so, what those differences are. Empirical evidence will be acquired by experimentation. Since submitting the project proposal (see Appendix A), changes have been made to the data collection methods—there is no longer an online questionnaire, to collect data of recommender systems most used by those who would have responded. In addition, the experiment proposed would have been more qualitative in nature, whereas now, collecting quantitative data is the main focus. Though, before there can be data collection, an extensive review of relevant literature will need to occur, in order to identify the key psychological constructs that are responsible for reported gender differences in the use of computer software, and more specifically, in intelligent interactive systems, including recommender systems. Then, to gain a clear understanding of the subject matter and to inform the empirical research stage—for each of the main psychological constructs identified, synthesize the following details from relevant studies: type of system, task being performed, interactions involved, measurement of gender difference, how gender differences were assigned to specific constructs, actual gender differences that were found, and the relevance of findings.

1.3 Overall research aim and objectives

The overall aim of this research is to investigate whether gender differences exist in the use of recommender systems—which are a type of intelligent interactive system. If differences are identified, understand why they may be happening and what this means to the design of these systems. Though, in order to achieve this outcome, it will be necessary to explore the subject matter and its context—performing a review of relevant literature—to use this

information to help formulate an experiment to collect enough data to explore, that once analysed, would be capable of providing empirical evidence to meet the overall research aim.

This can be broken down into the following four objectives:

1. Identify current theory regarding known gender differences applicable to the use of intelligent interactive systems.
2. Evaluate critically the gender differences identified, including details on the tasks and interactions being performed.
3. Investigate how participants interact with a prototype recommender system, including their reasoning for specific actions.
4. Discover, through the analysis of data collected, whether gender differences exist, describing any that are found.

1.4 Methods and work plan

Use the knowledge acquired from the literature review, including how other experiments identified gender differences, to design a between-participants experiment.

Recruit an equal number of male and female participants using convenience sampling—utilising social media to reach as many people as possible. Screening participants to ensure suitability.

Conduct the experimental study, collecting both qualitative and quantitative data. These will include timings, frequencies, transcripts from retrospective think-aloud sessions—as well as data from questionnaires (e.g. self-efficacy and perceived workload).

Analyse the empirical research quantitative data using statistical tests that will show whether a) there are gender differences, and b) the results are statistically significant. Utilising the qualitative data to provide participants' quotes, to clarify their motivations for specific actions (e.g. understanding why they added or removed a specific programme, or scrolled a particular row of programmes).

Present and discuss the findings, drawing conclusions as to whether gender differences were found, and if so, what they were—including the implications of results.

1.5 Value of this research

The overall aim of this research is to investigate whether gender differences exist in the use of recommender systems, and if so, what those differences are. Therefore, the beneficiaries would include software developers—for if there are differences found—these could have implications for the design of other systems that involve some kind of uncertainty. With this newly gained understanding, developers could make changes to their applications to better meet the needs of both males and females. Which in turn, could impact use and thus sales (for commercial software). Obviously, this would also impact the users of these systems, allowing for the types of interactions suited to both genders, rather than just one.

The results of this research would of course provide additional evidence (either in support or against the existence of gender differences) for the gender HCI domain, adding to the existing knowledge—which could be useful to other academics researching this, or related fields.

1.6 Outline structure

Chapter 1: Introduction and objectives

Providing background information about recommender systems, and why there may be gender differences in their use. Why the research is worthy of undertaking, and what that undertaking involves.

Chapter 2: Context

Delving into the constructs that explain the differences males and females experience in their software interaction. Understanding the experiments that have shed light on this issue, and finding out how a gender gap could be narrowed.

Chapter 3: Methods

A detailed breakdown of what is required to plan, conduct, record, analyse, and provide conclusions for an exploratory experiment to see if there are indeed gender differences in the use of a recommender system.

Chapter 4: Results and discussion

An exploration through the data collected, calling on the knowledge gained from the review of literature, to guide the direction of analysis, in search of data

with significance. Using both quantitative and qualitative techniques, to analyse the detailed interactions the participants would have with the recommender system.

Chapter 5: Evaluation, reflection and conclusions

Looking back over the various stages of the research; summarising the journey taken, while giving reflection on the experience—before concluding on the main findings discovered.

References

An alphabetical listing of the sources referred to in this report are included, using the Harvard referencing style (author-date).

Appendices

Appendices are at the end of this report, and are referenced throughout. Additional appendix materials that are in a format not suitable for this report (e.g. video recordings), will be provided on a separate DVD.

Chapter 2: Literature review

2.1 Introduction

This literature review focuses on objectives 1 and 2, as set out in sub-section 1.3 of Chapter 1 (Objectives 3 and 4 will be met through empirical research, which will be informed by the findings of this literature review):

1. **Identify current theory regarding known gender differences applicable to the use of intelligent interactive systems.**
2. **Evaluate critically the gender differences identified, including details on the tasks and interactions being performed.**
3. Investigate how participants interact with a prototype recommender system, including their reasoning for specific actions.
4. Discover, through the analysis of data collected, whether gender differences exist, describing any that are found.

Firstly, there is a brief discussion on exactly what is meant by both recommender systems and gender. Secondly, the key psychological constructs are identified—which are theorised to cause the gender differences reported, including: information processing theory, risk aversion, and self-efficacy; plus, further related topics (e.g. tinkering, mental models). Thirdly, emerging issues will be addressed, with an explanation as to why further empirical research is necessary.

2.2 Recommender systems

Recommender systems are considered a type of intelligent interactive system—the intelligent part of the names comes from the fact that machine-learning is involved (e.g. using algorithms that can learn from users' data to make predictions). Recommender systems (also known as recommender engines) have grown in popularity over the past 20 years, to the point that they are now ubiquitous. Examples can be found on some of the most popular websites (Wikipedia, 2015), from companies such as Amazon, Apple, Facebook, Google, Netflix, Twitter, Wikipedia and YouTube—they all use recommender systems (see Figure 1). These types of systems provide recommendations for all manner

of things, such as what to watch or read, who to be friends with, who to follow, what to buy.

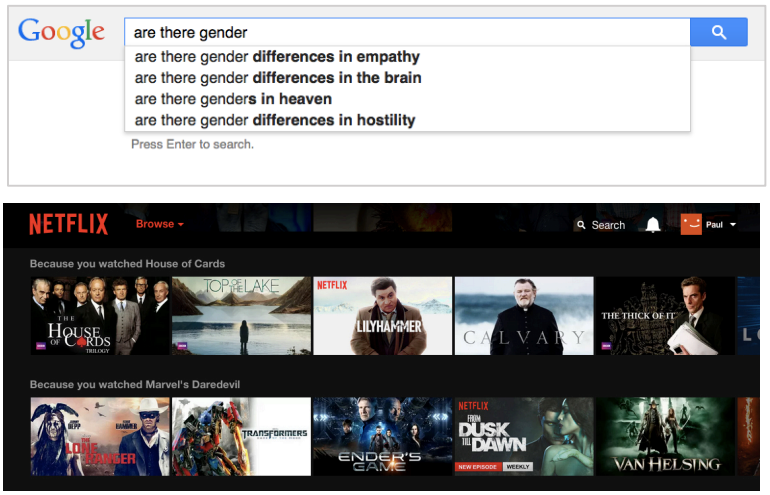


Figure 1. Examples of recommender systems on two popular websites: Google autocomplete; and Netflix 'Because you watched...' recommendations.

The way in which these recommender systems typically work, is by using one of two methods: collaborative or content-based filtering (or a combination of the two) (Melville and Sindhwani 2010; Ricci *et al.*, 2011). Collaborative-filtering is where past behaviours of users, are analysed in order to predict other items they may be interested in (e.g. because you read books A and B, book C will be recommended because other users who read books A and B also read book C). Alternatively, content-based filtering uses specific aspects about an item to provide recommendations for other items that share those attributes (e.g. because you watched comedy films previously, other comedy films are recommended).

Recommender systems are usually considered to be low-cost, due to the sorts of recommendations they provide (e.g. books, movies, music, people to follow)—which are unlikely to involve decisions that would have a huge impact on things which are high-cost, such as health or finance (Ricci, 2011; Bunt, Lount and Lauzon, 2012). The study of recommender systems as an independent area of research began over 20 years ago (Ricci, 2011).

2.3 Gender

Gender is defined by Oxford Dictionaries (2015) as: "the state of being male or female (typically used with reference to social and cultural differences rather than biological ones)". A crucial distinction is made here between sex, which is

biological, and gender, which is psychological. This theory of gender being a social/cultural construct, rather than purely biological echoed by World Health Organization (2015), when they say:

> "Gender refers to the socially constructed characteristics of women and men – such as the norms, roles and relationships that exist between them. Gender expectations vary between cultures and can change over time. It is also important to recognize identities that do not fit into the binary male or female sex categories."

Regardless of how much of this construct we call gender is down to biology versus societal influences, the fact remains that there are reported differences between how males and females use software—as will be discussed in the remainder of this chapter. In Beckwith *et al.*, (2006a), a new term was used for this relationship between gender and human-computer interaction (HCI), called Gender HCI.

2.4 Psychological constructs (gender differences)

A psychological construct (also known as a hypothetical construct or construct) is a scientific theory that is used to describe an aspect of human behaviour that can not be observed directly (Encyclopedia Britannica, 2015). Common in HCI are several such constructs, one example being risk aversion. Risk aversion can not be observed, however, risk averse behaviour can (e.g. not performing an action due to uncertainty of the outcome). Therefore, to conduct research into who are more risk averse, males or females. Researchers could observe both groups, taking measures of risk averse behaviour. Comparing these measures to see which group had exhibited the highest number of risk averse behaviours—thus considered more risk averse. This of course it a simplistic explanation, though does provide a basis for discussions later in this chapter.

Next, the main constructs for which gender differences have been identified in relevant literature will be discussed one by one, examining the research that has currently taken place, to discover emerging issues and where further empirical research is needed.

2.5 Information processing theory

The theory behind information processing in relation to gender, is that females favour a comprehensive style (e.g. acquiring a breath of information before acting), whereas males prefer the use of selective styles (e.g. acting on the first piece of information). As discussed by Meyers-Levy (1989) where they found that females were comprehensive in their informational needs from the outset, whereas it was only once problems become difficult that males looked for this information too. Which was also reported in another study, that found females will not take action until they have a lot of information (Beckwith *et al.*, 2005b). Arcand and Nantel (2012) also found concurring evidence for the information processing theory, and in addition, they go on to say in relation to women, that they "elaborate more on the information presented, tending to make more associations between the various pieces of information, comparing brands based on various attributes, or linking products to contextual information.".

Beckwith *et al.* (2006a) found in one of their studies, that gender differences in problem solving spreadsheet software, with information processing styles being one of the factors. They go on to suggest that as no one male, or no one female will process every trait assigned to a particular gender, then to design software to support information processing differences, will in fact benefit all users, male and female. A fine sentiment, though how to achieve this, they do not surmise.

Having time to process information was found useful in research by Beckwith *et al.* (2006b), where they identified that females paused following an action, more often than males. With pausing benefitting both males and females by providing extra time to reflect on how the system responded—increasing the users understanding of the debugging functionality and their own use of it. Of course, being able to gain a better understanding of how a system works which appears a logical choice, as Grigoreanu *et al.*, (2008) found to be the case when users were debugging a system, it was found that information about strategy emerged as a primary requirement. However, explanatory information is not always necessary, Bunt, Lount and Lauzon, (2012) found that in many deployed recommender systems which lack explanations for how they provide their recommendations, or are simply not clear—"most users understood the general idea behind the systems, but lacked knowledge on the details". Suggesting their existing mental models for how the system might work, were intact. In addition, they found that "participants doubted that they could personally benefit from viewing an explanation.". Unfortunately, gender was not treated as a variable in

that particular piece of research, which is a shame as they had equal numbers of males and females for one of the experiments, which would have been interesting to see if any gender differences existed, and of any significance.

2.6 Risk aversion

Risk aversion deals with the idea of an unknown outcome, where there is the fear that an action could produce an unfavourable result, for which the individual involved would rather avoid (Kahneman and Tversky, 2000). A person's gender has been found to effect risk aversion and the perception of risk, this can also be tied to information processing, as females will need more details than males before they are willing to take action (Beckwith *et al.* 2005b). Therefore, providing females with more information could aid understanding and thus alleviate the perception of risk. Kulesza, *et al.* (2012), discuss how mental models can help the user better understand the systems they are interacting with, and highlight the two varieties of mental models: functional (shallow) models, and structural (deep) models. The difference between the two being that functional models assumes higher-level understanding of the system, just not detail. Whereas, structural models give details for how and why it works, the way it does.

As with many of the themes addressed in this chapter, there is much overlap between the various psychological constructs and how they manifest themselves in the behaviours of end-users who user the software. Risk aversion, self-efficacy, tinkering, information processing can all influence each other—which leaves the researcher to address the role of each in the particular action, relying on theory to provide answers that may not always be clearly defined.

2.7 Self-efficacy

Self-efficacy is the strength of belief a person has in themselves to be able to complete a certain action or task (Bandura, 1986). Originating from the field of social psychology, self-efficacy is frequently studied in relation to the use of computers and computer software. Ways in which to measure self-efficacy in HCI have been devised, with a popular example being a validated self-efficacy questionnaire by Compeau and Higgins (1995). The questionnaire consists of ten task-focused questions—designed to measure one's confidence in using a hypothetical (and unfamiliar) software application.

One of the first people to report a connection between self-efficacy and gender was Busch (1995)—who found that female students on a computer course he taught—had significantly lower self-efficacy than male students in performing complex spreadsheet tasks. Since then, self-efficacy has been a common measure when investigating the existence of gender differences in computer-related tasks.

The role of debugging machine-learned programs, which includes intelligent interactive systems, can involve users being able to 'correct' learned behaviour of the system. Kulesza *et al.*, (2010) rationalise the need for this functionality to exist, as many machine-learning algorithms learn behaviour rules from individual users, creating a 'program' that instructs the system in how to respond to future inputs. Thus, they say, if mistakes are made by the program, the user needs a way of correcting them. Research involving users debugging systems, have found gender differences in relation to self-efficacy. Beckwith *et al.* (2005a), found that there were differences between genders, finding that "females had lower self-efficacy than males did about their abilities to debug" and "females were less likely than males were to accept the new debugging features". Which has implications considering that Grigoreanu *et al.* (2008) found that self-judgement (a closely related construct to self-efficacy) is a significant factor when attempting to 'fix' a system.

Self-efficacy is used as a variable in several of the studies discussed, often used along gender in tasks involving debugging. There is another gender difference that arises in several studies (Beckwith et al., 2007; Grigoreanu *et al.*, 2008), which has also been measured along with gender and self-efficacy—and that is the concept of tinkering. As the name suggests, tinkering is the playful experimentation a user can perform when using software (Rowe, 1978). A study by Beckwith *et al.* (2006b), set out to investigate "whether gender, tinkering, and self-efficacy interact to predict effective use and understanding in user debugging". What they found, surprisingly, was that although males did tinker more than females, it was often detrimental to how effective they were in debugging the system. This was unexpected, as earlier work on computer-related gender differences had found the behaviour of males to align with problem-solving required in user programming.

Beckwith *et al.*, (2005a) advocates that designers of software products make appropriate accommodations for gender differences, or else it could affect how willing a group of users are to accepting new features that could benefit them. The design of a system can create barriers, which have been found to affect

one gender more than another. As is the case for Kulesza *et al.* (2009) who found in their study involving debugging machine-learned email filing, that there were gender differences in the number and sequence in barriers. In the Grigoreanu *et al.*, (2008) paper however, one aspect worth noting is that although gender differences were observed (in debugging spreadsheets), through feature design they were able to demonstrate a closing of the gender gap. Another of their findings was that as a result of changes to the system, females were more interested in debugging, and tinkered with elements of the system—something the females in the control group without the feature design change did less. Even though just one example, it is encouraging that design changes can help bring down the barriers dividing the genders.

2.8 Emerging issues and the need for empirical research

Before going any further, it should be mentioned that although research exists that does find gender differences—as mentioned throughout this chapter—generalisations should be avoided. This is discussed by Nelson (2012), suggesting "The statement 'women are more risk averse than men' is fundamentally a metaphysical assertion about unobservable essences or characteristics, and therefore cannot be empirically proven or disproven." before going on to say "The widespread acceptance of such statements appears to perhaps be rooted more in confirmation bias than in reality.".

Certain behaviours could occur for many different reason. It should therefore be clear when research is reported, why the particular behaviour is found to be indicative of the construct that it is being reported as. And that plausible alternative theories for the actions, are ruled out. In saying that, many of the studies did provide qualitative data in the form of what participants had said while performing an action, to provide their motivation for that action (Bunt, Lount and Lauzon, 2012; Kulesza *et al.*, 2015). This does appear to be a very successful way of providing evidence for claims of a particular construct. And therefore, definitely worth undertaking when collecting empirical data.

Spreadsheets are a popular software it seems for experiments involving gender differences. They allow for many of the interactions that facilitate finding differences, such as information processing, debugging, tinkering, risk. It makes sense if the study is interested in how males and females interact with software differently, to create a study that is likely to incite the desired outcome.

Alternatively, it could just be the case that the same software (or similar) is used across a number of studies to allow for data comparison.

There is also the issue of the system being capable of providing enough data points to analyse. Thinking specifically about recommender systems of the type being investigated for this project, there could definitely be more research undertaken. Though devising experiments to get useful data could be a challenge. However, after reviewing the experiments and the tasks participants were instructed to undertaken across the research scrutinised for this literature review, there are opportunities with the right system to ensure for successful and useful data collection. Devising tasks that would require participants to behave in ways to signify the specific constructs being investigated (i.e. gender differences)—ensuring there are a range of interactions open to the user, that can be measured in order to make comparisons between males and females. Whether in the form the frequency of a particular action is performed, or the time in takes to perform an action.

Another finding derived from reading the broad range of literature across this topic is how much mental models play a role in participants' understanding of how the various intelligent interactive systems work to produce their particular feedback (e.g. recommendations). Kulesza, et al. (2012) found that it was the participants who were able to improve their mental models of how the system worked throughout the study, were the ones who benefited most by gaining knowledge to help them better use the system to their own satisfaction. Therefore, they advocate the intelligent systems are designed to provide more feedback, and of a higher quality, to assist the user to build a more accurate mental model.

For these reasons outlined above, there is scope for and a benefit to conducting research into the existence of gender differences in the operation of recommender systems. The areas identified that would be useful to explore include information processing and risk aversion, as well as the impact self-efficacy may have on them. If differences were found, and they were significant, then it would have implications for the design of other systems that involve some kind of uncertainty—particularly those involving low-risk decisions.

The next stage of this research will document the research methods used to obtain the empirical data, including details on the research strategy, participant selection, data collection methods, and materials used.

Chapter 3: Methods

The literature review, which was discussed in the previous chapter met objectives 1 and 2, allowing objective 3 to be undertaken—the process of which is documented fully in this chapter (Objective 4 will be met through analysis of the empirical study):

1. Identify current theory regarding known gender differences applicable to the use of intelligent interactive systems.

2. Evaluate critically the gender differences identified, including details on the tasks and interactions being performed.

3. **Investigate how participants interact with a prototype recommender system, including their reasoning for specific actions.**

4. Discover, through the analysis of data collected, whether gender differences exist, describing any that are found.

An empirical study was conducted, using a between-subjects design. Both quantitative and qualitative types of data were collected. Participants undertook three tasks using a prototype recommender system. Prior to the tasks, participants completed a self-efficacy questionnaire, and after each of the three tasks completed a NASA-TLX questionnaire. Sessions concluded with a retrospective think-aloud covering all three tasks, ending with a semi-structured interview. Data collected were analysed using both quantitative and qualitative methods.

3.1 Participants

3.1.1 Eligibility

All participants were over the age of 18 years old and not considered vulnerable. An equal number of male and female participants were necessary due to the nature of the research. Screening was used to ensure no participant had a computer-science background which involved recommender systems (i.e. that they would have a technical understanding of how these systems work).

3.1.2 Recruitment

Convenience sampling was used for finding participants. Firstly, an email recruitment advert was used as part of the strategy to find participants (see

Appendix B). The email was sent to HCI students at City University London, who were encouraged to forward the email on to people that they know who may be suitable to take part.

Secondly, participation requests were made via social media, with Twitter and Facebook used—posted via the researcher's personal and business accounts, as well as two unofficial City University London Human-Centred Systems closed Facebook groups. With all postings again asking for these requests for participants to be shared to the viewer's own networks, in order to reach a wider population of people. These various means of advertising combining to reach a potential audience of over 600 individuals—this is before sharing took place, which would have increased that number further.

Interested participants were requested to respond by email with their details. This allowed contact to be made via email or telephone to screen for suitability, using a recruitment screener (see Appendix C). Participants considered suitable were then able to select a suitable timeslot. Up to eight timeslots were available each day, over a two-week period (17th-28th August 2015)—with the earliest available week-one timeslots offered first, to make the best use of time available.

An hour was allotted for each session, with a 30-minute gap between each one—allowing for slight overrun or participants arriving late. This also gave time to reset the software, and perform other duties such as data backup (see Appendix D for post-session activities).

3.1.3 Participant numbers

Due to the timescale for this research and its duration, not only to collect data, but to process and analyse it—the maximum number of participants considered to be achievable was 20 (10 males and 10 females).

3.2 Experiment design

3.2.1 Between-subject experiment

A between-subject experiment was selected as a means to collect the necessary data, to allow for potential gender differences to be observed between the two groups of participants (males and females).

3.2.2 Independent variable

The independent variable in this experiment, manipulated by the experimenter, is gender (i.e. comparing males and females).

3.2.3 Dependent variables

The dependent variables, assumed to be dependent upon the independent variables, and thus through this experiment either established or dismissed, are:

- Self-efficacy – obtained using Compeau & Higgins questionnaire
- Perceived workload – obtained using NASA-TLX questionnaire
- Debugging*
- Information processing theory*
- Risk aversion*
- Tinkering*

The literature review in the previous chapter provided the types of gender differences that are known to exist in software use (e.g. self-efficacy, debugging, information processing theory, risk aversion, tinkering), with examples of the experiments and systems used (e.g. Grigoreanu *et al.*, 2008; Kulesza et al., 2009; Bunt, Lount and Lauzon, 2012).

*Measures of frequency are used in the analysis of these variables, and will be discussed later in this chapter, under analysis.

3.2.4 Experimental requirements of system

This research set out to investigate how males and females use a recommender system, to see if differences exist. It was therefore important to use a system permitting interaction that would meet three requirements:

1. Allow for a feedback loop to occur between the participant and the system (i.e. when the participant performs an action, the system makes a change as a result of that action, which is perceivable to the participant).
2. Be simple enough so that with minimum instruction a range of actions could be performed.
3. The actions that the system allows the participant to perform, allow differences (if they exist), to be observable.

Thus, a recommender system was sought that would offer recommended items, which the participants could provide feedback on, in a manner that was simple to perform.

3.3 Prototype recommender system

A search to find a suitable recommender system was carried out, that involved considering prototypes used from previous studies—if said prototypes were freely accessible. Alternatively, searches were performed via the internet for existing software that would meet the three requirements listed previously. Websites such as Google.com, Github.com, and Stackoverflow.com were searched using appropriate terms (e.g. 'recommender system resources', 'free recommender system'). This second route turned out to be successful, resulting in a prototype recommender system from BBC R&D being found, called Sibyl Recommender System (Sibyl.prototyping.bbc.co.uk, 2015).

Sibyl (the Greek word for a female prophet) is a web browser-based system that provides recommendations based on feedback (familiar programmes) provided by the end-user using a drag-and-drop user interface. The system tries to predict which programmes the end-user is likely to watch, based on programme-to-programme similarity (Sibyl.prototyping.bbc.co.uk, 2015).

The way in which the system selects which programme it recommends is described as follows:

> "In the collaborative filtering approach the programme-to-programme similarity is determined by counting the number of common viewers for each pair of programmes. In effect, the approach can be described as "people who watched this programme also watched..." and tends to result in a wider variety of recommendations. Sometimes the results can appear eerily prescient whilst at other times it can be predictably populist." (Bbc.co.uk, 2015)

The data the system uses to provide its recommendations is updated frequently. It was therefore crucial for the study that these data would not change between sessions, ensuring each participant had the same programmes provided by the system. As these data are contained in JavaScript files, which are stored front-end, these could be downloaded to a computer. This allowed a fully-functioning instance of the system to be downloaded (obtained on 08.08.15), that could be run independently of internet access.

Specific changes were then made to the system, both visually and in functionality, in order to make the system suitable for the research purposes (see Figure 2). Visual changes included removing text that could be distracting to participants, making the title style identical of each of the three boxes (i.e. Recommendations, Like, Dislike). Alterations to its functionality included removing controls to select content by All, TV, or Radio. This was to ensure participants could not switch content type. Two genres were also removed, Children's and Learning, as these were not considered appropriate for the participants' ages. In addition, those two genres were not displayed when selecting the All category (all other genres were). A final change made, was to disable the programmes' thumbnail/title from linking to iPlayer—as this was not necessary.

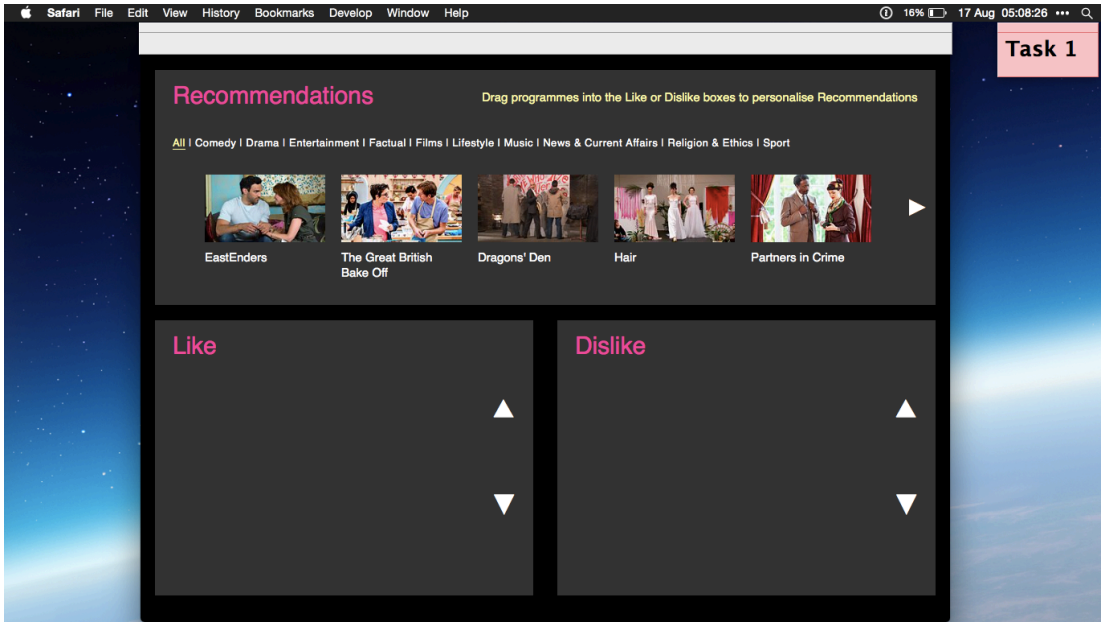


Figure 2. Screenshot showing the recommender (adapted from the BBC's Sibyl Recommender System).

See Appendix E for Task Setup Screenshots.

The recommender system consists of three main elements (which are all visible on screen without the need to scroll the actual browser window):

1. Recommendations. At the top of the browser window is a box displaying recommended programmes in a horizontally scrollable list (five programmes are visible at any one time, with most recommended programme on the left). The All category (default) contains all 294 available TV programmes (346 for the radio version). Above the

programmes, is a genre menu—featuring a row of clickable buttons that filter the programmes by genres.

2. Like. In the bottom left of the window is a box for the liked programmes to be dragged to/from, which can be scrolled vertically (three programmes are visible at any one time).
3. Dislike. In the bottom right of the window is a box for the disliked programmes to be dragged to/from, which can be scrolled vertically (three programmes are visible at any one time).

Programmes can be moved between any of the three boxes (Recommendations, Like, Dislike). This action provides the system with new information about the end-users' preferences, resulting in the recommended programmes row (in the Recommendation box), updating as a result of what is currently contained within the Like and Dislike boxes. For example, if starting with no programmes in the Like or Dislike boxes, the action of moving Eastenders from the Recommendation box into the Like box, would cause the row of recommended programmes to update, displaying similar programmes, such as other soap operas/dramas of a similar genre.

See DVD for a copy of the prototype used in this study.

3.4 Sessions

3.4.1 Location and equipment

As the main piece of equipment necessary for the experiment sessions was a laptop containing the recommender system and recording software, the benefits of a strict lab-based setting was not considered crucial. Instead, allowing sessions to be conducted at locations convenient for the participants was seen as more important—especially considering the short time frame for recruitment and data collection. Therefore, locations that were safe, quiet and relaxed will be used, either at City University London or elsewhere.

The following equipment and software was used for all sessions:

- 11.6" MacBook Air (laptop), running OS X Yosemite (version 10.10.5)
- Wired two-buttoned mouse
- Two versions of the adapted recommender system (TV and Radio)
- Safari browser (version 8.0.7)

- Mac application called Stickies
- QuickTime Player (version 10.4)

The only equipment that did vary, depending on location, were the tables and chairs used—though not considerably (i.e. all tables/desks used were of a similar height; with chairs appropriate for the task—allowing arms to move freely).

3.4.2 Session length and dates

Each session lasted for up to one hour, for which the pilot session provided timings. All twenty sessions took place over a two-week period (17th-28th August 2015), with the pilot taking place on the 13th August 2015.

3.4.3 Session sequence

A consistent order of actions was used for each session, as follows:

1. Perform pre-session checks (e.g. ensure systems were reset)
2. **Participant arrives** – welcome and introduction
3. Give information sheet to participant and consent form to complete
4. Participant completes background / self-efficacy questionnaire
5. Read instructions – Practice task
6. Start recording
7. Read scenario and instructions – Task 1: Own Preference
8. Participant completes NASA-TLX questionnaire for Task 1
9. Read scenario and instructions – Task 2: Box Limit
10. Participant completes NASA-TLX questionnaire for Task 2
11. Read scenario and instructions – Task 3: Add Only
12. Participant completes NASA-TLX questionnaire for Task 3
13. Stop recording and save file to desktop
14. Start a second recording and play the first recording
15. Participant performs retrospective think-aloud
16. Participant interview
17. Stop second recording and save to desktop
18. **Participant leaves** – after thanking them
19. Perform post-session tasks (e.g. backup data, reset systems)

3.4.4 Computer setup

Each session requires the participant to perform a practice task followed by three main tasks. To avoid having to set the recommender system up for each different task, it is instead easier to have four separate desktop Spaces pre-setup on the computer. These Spaces are simply additional desktops in OS X that can be easily switched between.

The mouse was placed to the right of the computer. Thought after arrival participants were asked if they wanted it moved to the left. Participants were instructed to use the mouse only (not the trackpad), to ensure consistency between participants.

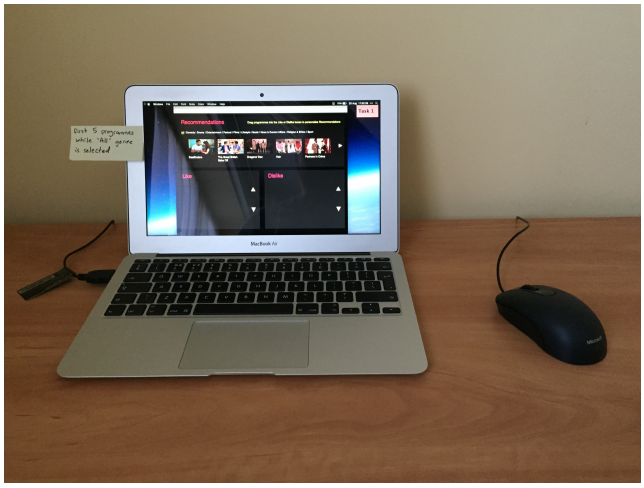


Figure 3. Setup showing laptop computer and mouse on a desk.

3.4.5 Recordings

Main tasks were video-recorded using QuickTime, which captured the on-screen activity, along with audio. Video of the participants' faces was not recorded, as capturing timings were the focus rather than expressions useful during a think-aloud. Another reason for not capturing participants' faces, was that for the retrospective think-aloud later in the session, participants would be required to watch back the recording of their tasks—having their face visible on screen during this could have been distracting.

After each session, recorded data were backed up to a flash-drive—transferred to a second computer that evening, and stored securely.

3.4.6 Practice task

To avoid learning effects, the practice session involved radio programmes, rather than the TV programmes used for the main experiment tasks. This

required a second version of the recommender system to be created, which was identical to the first, except for the type of content (radio instead of TV).

Participants were given a brief overview of the system and then allowed to use it for two minutes, so they could practice interacting with it.

See Appendix F for Session Script.

3.4.7 Main tasks

There were three main tasks that the participants were asked to complete. Participants were given a printed Instruction sheet before each task, which was a copy of the instructions readout to participants by the facilitator. Each participant completed three tasks, in the following order:

1. **Own Preference.** In which participants used the software to achieve five recommended programmes in the All category, that they themselves would want to watch.

2. **Box Limit.** In which participants used the software, with the Like and Dislike boxes restricted to each containing no more than three programmes at any one time—in order to achieve five recommended programmes in the All category—for a fictional friend who loves history programmes.

3. **Add Only.** Participants used the software, with the restriction of not removing any programmes from the Like or Dislike boxes—in order to achieve five recommended programmes in the All category—for a fictional friend who hates sport, history and nature programmes.

As each participant was required to complete all three tasks in the same order—along with the fact that the participants were the independent variable, not the tasks—this meant that order-effect bias was not a factor that required consideration.

3.4.8 Task completion

As there was no objective way of telling whether participants had completed a task (i.e. it was down to their own judgement of programme suitability), it was therefore up to participants to verbally inform the facilitator when they were finished.

3.4.9 Task success

Tasks were considered to be successfully completed if the participant had followed the instructions for that tasks. However, there was a small amount of flexibility here, which was based on judgement of whether there would be little or no effect to overall frequencies. And to ensure consistency, flexibility given to one situation, was applied to all occurrences of that same type throughout the tasks.

These flexibility criteria are best demonstrated with two examples. Firstly, if in the Box Limit task (T3) a participants did not restrict the Like box to containing only three programmes at any one time, instead adding as many as they wished. Then this could obviously greatly effect the mean average for how many programmes were added to the Like and Dislike boxes (as well as impacting other frequency counts), and so would be classed as an unsuccessful task. A second example using the same task, would be where the participant added four programmes to the Like box, then realising their mistake, thus corrected this by removing a programme, to leave just three. In this second situation, the task would be considered successful.

3.4.10 Retrospective think-aloud protocol

Following the tasks, participants were shown the recording of their on-screen interaction with the system and asked to perform a retrospective think-aloud. Asking them to discuss their reasoning for the actions they made and how they were expecting the system to react to the choices they made.

Participants were informed that they could pause the video at any point, and as often as they need to, so that they could fully explain their reasoning for every action they made.

3.4.11 Semi-structured interview

After the retrospective think-aloud, participants were asked a few final questions in order to assess how accurate their mental-model was for how the recommender system functions and produces its recommendations.

1. Please explain to me how you think the recommender system actually works?
2. When you drag a programme into the Like box, what do you think the system does ‘under the hood’ that causes what it recommends to change?

3. What do you think would happen if the programme Mastermind was dragged into the Like box? [confirm they know what the show is]
4. In hindsight, was there anything you could have done differently to improve the recommendations the system provided?
5. When you first used the system, before any programmes were in the Like or Dislike box, how do you think the recommendations at the top were ordered?
6. Finally, is there anything further you would like to add about your use with the recommender system?
7. See Appendix G for Interview Questions.

3.5 Materials

3.5.1 Session checklist and task setup instructions

A checklist of things that needed doing after each session was created, which included data backup and resetting the four browsers with the different recommender system setups.

See Appendix H for Researcher Instructions.

3.5.2 Information sheet and consent form

Participants were given an information sheet that provided details about the research project and what their involvement would be. Participants were then given a consent form to complete and sign—giving informed consent for the collection and use of data, such as video-recordings.

See Appendix I and J for Information Sheet and Consent Form Template (see DVD for Signed Consent Forms).

3.5.3 Researcher instructions (session script)

A script was devised that allowed the facilitator to follow throughout the session, which ensured consistency across sessions. This included all instructions and task details, and prompts for important actions—such as starting the recordings.

See Appendix F for Session Script.

3.5.4 Background questions

Only the gender and age of the participants were asked. These two questions were added to the start of the self-efficacy questionnaire. The participants had already been screened for suitability, which meant there was no need to repeat questions from the screener.

3.5.5 Self-efficacy questionnaire (Compeau & Higgins)

At the start of each session, participants answered a widely used, validated self-efficacy questionnaire (Compeau and Higgins, 1995) to measure their confidence in using a hypothetical (and unfamiliar) software application.

See Appendix K for Self-efficacy questionnaire template.

3.5.6 Perceived workload questionnaire (NASA-TLX)

After each of the three tasks participants completed a NASA-TLX survey (Hart and Staveland, 1998) to measure perceived task loads on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales. Which allows for ratings from 0-100, in increments of 5, with marks by participants between two ticks rounded up.

See Appendix L for NASA-TLX questionnaire template.

3.5.7 Participant instructions

Participants were given a printed Instruction sheet before each task, which was a copy of the instructions readout to participants by the facilitator. This was to reduce their cognitive-load of having to remember all of the instructions.

See Appendix M for Participant Instruction Sheet.

3.6 Pilot session

A pilot session was conducted four days prior to the first scheduled session. The pilot was used to test every element of the session would run as intended, to do this the pilot was conducted exactly as the actual sessions were intended to be run. This included: completing questionnaires, performing the practice task, the three main tasks, a retrospective think-aloud, and interview.

As a result of the pilot, a change was made to task 2, limiting the amount of programmes contained in either the Like or Dislike box, to just three at any one time. This was to force participants to consider their choice of selected

programmes more carefully, as it was felt the pilot participant rushed during this task. Sticky notes were added to the bezel of the screen before tasks, to remind participants of instructions relevant to the current task (see Figure 4).

A final change was made to cover the browser menu bar with a digital sticky note (using a Mac application called Stickies). This was done to stop an undesirable action that happened twice in the pilot, where the participant refreshed the browser "out of habit" to reset the Like and Dislike boxes (instead of removing the programmes manually).

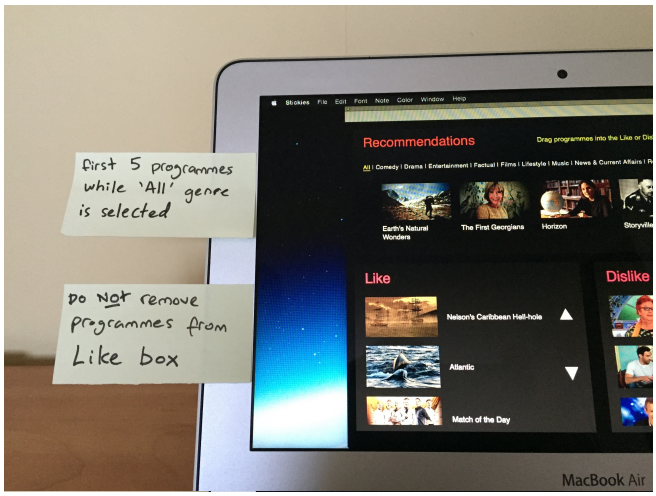


Figure 4. Sticky notes as reminders of task instructions.

3.7 Quantitative analysis

3.7.1 Independent variables

The independent variables in this experiment, manipulated by the experimenter, are males and females (i.e. grouped by gender).

3.7.2 Dependent variables

The dependent variables, assumed to be dependent upon the independent variables, and thus through this experiment either established or dismissed, are:

- Self-efficacy—obtained using Compeau & Higgins questionnaire
- Perceived workload—obtained using NASA-TLX questionnaire
- Debugging
- Information processing theory
- Risk aversion
- Tinkering

Using information from the literature review (e.g. known gender differences, and actions that signify said differences) will allow for investigation and analysis of the data obtained. For example, taking risk aversion, do females take longer to make actions to the system? And if they do, can it be attributed to risk aversion or information processing theory? (Are females in fact wanting a more comprehensive overview of information before making a decision?) This is where qualitative analysis of what participants said during the retrospective think-aloud helps to clarify intention.

3.7.3 Analysis process

Data from background, self-efficacy and NASA-TLX questionnaires were transferred into Excel.

See DVD for Spreadsheets containing all raw data.

Task frequencies sheets were used while watching each of the task video-recordings. This allowed easy noting down when specific actions occurred—which could be totalled at the end to provide action frequencies. The specific actions for which frequencies were recorded were:

- Time to first action
- First action (add, remove, switch, scroll, click)
- Time to add/remove first programme to/from Like/Dislike box
- When in All, number of times programmes were scrolled
- Number of different genres clicked (out of 10)
- When in a genre (excluding All), number of times programmes were scrolled
- Number of times genres were clicked (excluding All)
- Number of times All was clicked
- Number of programmes added to Like box
- Number of programmes removed from Like box
- Number of programmes added to Dislike box
- Number of programmes removed from Dislike box
- Number of pauses (of 3 seconds or more)
- Total task time
- Task Completed

- Like box scrolled
- Dislike box scrolled

Although the literature review provided the most reported types of gender differences and how these may appear in software use. It was still important to be thorough in collecting as many data points as possible, which could provide new opportunities of finding statistically significant differences between males and females when using the recommender system.

See Appendix N for Task Frequencies Sheet template.

Timings data were also added to the Excel spreadsheet, which allowed inspection of the data to occur to see if there were early indicators of gender differences (e.g. obvious contrasts in numbers when columns are sorted by gender).

IBM's SPSS Statistics software was used in performing all statistical analysis. Data was transferred from the Excel spreadsheet when performing specific tests.

Dancey and Reidy (2014) proved to be a great resource when performing statistical analysis—providing clear guidance which was useful for all tests performed. The general process was to check the data to see if it was normally distributed using histograms—which the majority of the data collected from this study was not, which is common when using small sample sizes, according to Dancey and Reidy. This meant non-parametric methods would be required—and as differences in between-participants' data was sought, this led to the use of Mann-Whitney *U*-tests. With other charts produced (e.g. box plots), and tests conducted as and when required (e.g. 2 x 2 chi-square).

3.7.4 Excluded data

If a task was considered to be unsuccessful (i.e. the participant did not follow the instructions), then certain data from that task would be excluded from analysis, where its inclusion would be inappropriate (e.g. if a participant had 30 programmes in the Like box, yet the task instructions stipulated only 3 at any one time could be in there. Then it would be inappropriate to use total task time, number of programmes added to Like box, whereas, time to first action could still be used).

3.8 Qualitative analysis

3.8.1 Transcripts

Full transcripts were produced from audio-recordings made during the retrospective think-aloud sessions and the interviews that followed. Quotes were taken from these data and used to reinforce the motives behind types of actions that resulted in quantitative data results. To enable the finding of quote, these data were coded by topics relating to the dependent variables for this study (e.g. self-efficacy, risk aversion, information processing theory).

See DVD for Spreadsheet of Full Transcripts.

Chapter 4: Results and discussion

The completion of objective 3, as documented in the previous chapter, allowed analysis of the empirical study (i.e. objective 4) to take place:

1. Identify current theory regarding known gender differences applicable to the use of intelligent interactive systems.
2. Evaluate critically the gender differences identified, including details on the tasks and interactions being performed.
3. Investigate how participants interact with a prototype recommender system, including their reasoning for specific actions.
4. **Discover, through the analysis of data collected, whether gender differences exist, describing any that are found.**

This chapter presents and discusses the results from the quantitative and qualitative analysis of the data gathered. The results are grouped into the following sections:

- Participants' background
- Self-efficacy
- Task completion and success
- Task load
- System use
- Results summary

4.1 Participants' background

4.1.1 Age and gender

The ages of the 20 participants ranged from 22 up to 64 years old, with the mean age for males being 35.7 years old and for females being 38 years old (see Table 1).

As the purpose of the study was to investigate whether there are gender differences, an equal number of male and female participants were recruited (10 males and 10 females).

	Male	Female
Mean	35.7	38
Median	34	37
Standard Deviation (SD)	10.53	14.97
Number of participants	10	10

Table 1. Participants' ages (mean, median and standard deviation), grouped by gender.

See DVD for Spreadsheet of Participants' Details.

4.2 Self-efficacy

4.2.1 Participant ratings

At the start of each session, participants answered a widely used, validated self-efficacy questionnaire (Compeau and Higgins, 1995) to measure their confidence in using a hypothetical (and unfamiliar) software application.

Results from the questionnaires show that self-efficacy ratings were similar between gender groups (see Figure 5), with a mean of 63.3 (SD 20.35) for male participants, and for females participants, a mean of 56.9 (SD 20.95).

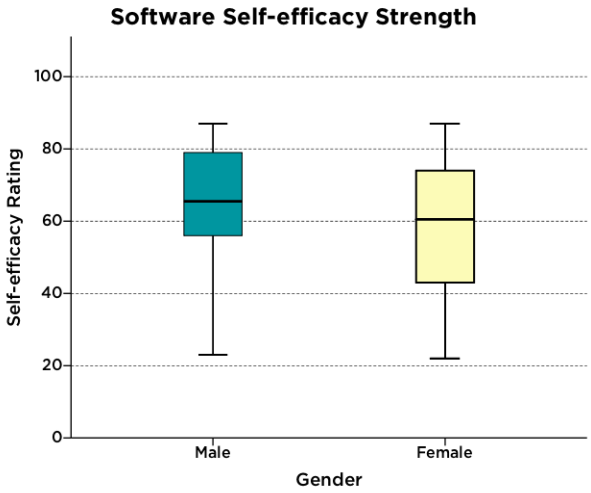


Figure 5. Box plot of participants' self-efficacy strength, grouped by gender.

This closeness in means, along with similar medians (males 65.5, females 60.5), suggests that participants, when grouped by gender, had a similar level of confidence towards using a hypothetical (and unfamiliar) software application. Thus, suggesting that any significant gender differences found in the data from the participants' use of the recommender system, is unlikely to be due to self-

efficacy alone. These scores are quite unusual as in most of the reported studies, as evident in the literature review, females score consistently lower than males. This then has implications in how they interact with the system—more tinkering, or a requirement, as mentioned earlier, that females will want more information before they make changes to the system Therefore, the fact that these scores are this close could impact the findings throughout the study, as there is not the usual divide to begin with.

See DVD for Spreadsheet of Self-Efficacy Questionnaire Data.

4.3 Task completion and success

4.3.1 Task completion

All tasks were completed by the 20 participants.

4.3.2 Task success

Only two out of the 60 tasks were not successful. These were both Box Limit tasks (T2), where two different participants (P6, female; P11, male) failed to follow the instructions to not have more than three programmes in either the Like or Dislike box at any one time. This was due to misunderstanding the instructions (P6), or simply forgetting (P11).

See DVD for Session Videos (including tasks, retrospect think-alouds, and interviews).

4.4 Task load

4.4.1 Perceived workload

After each of the three tasks participants completed a NASA-TLX survey (Hart and Staveland, 1998) to measure their perceived workload on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales. Allowing ratings from 0-100, in increments of 5—with marks made by participants, if falling between two ticks, rounded up.

For each of the three tasks, female participants recorded a higher perceived Task Load Index (mean) rating than male participants (see Figure 6 and Table 2). However, Mann-Whitney tests for each of the three tasks show these differences to be statistically insignificant (see Table 3).

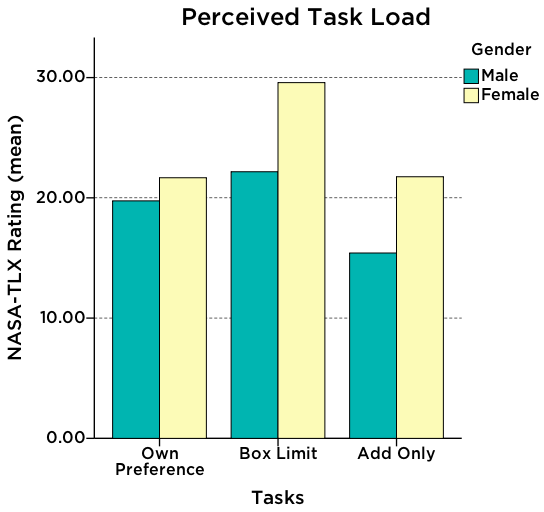


Figure 6. Participants' perceived task load rating (mean), grouped by gender, for each of the three tasks.

	T1: Own Preference		T2: Box Limit		T3: Add Only	
	Male	Female	Male	Female	Male	Female
Mean	19.57	21.67	22.17	29.59	15.42	21.75
Std. Dev.	12.29	12.58	14.19	17.21	8.33	11.13

Table 2. Participants' perceived task load rating (mean and standard deviation), grouped by gender, for each of the three tasks.

	T1: Own Preference	T2: Box Limit	T3: Add Only
Median (gender)	17.5 (m), 21.25 (f)	16.67 (m), 30.84 (f)	15.42 (m), 21.25 (f)
Mann-Whitney *U*	46.5	37	30
z	−0.265	−0.984	−1.512
p-value (exact sig. 2-tailed)	0.81	0.342	0.138

Table 3. Mann-Whitney tests found no statistical significance for females having higher perceived task load ratings than males across all three tasks.

See DVD for Spreadsheet of Perceived Task Load Data.

4.5 System use

4.5.1 Time to first action

As each task was different and new restrictions were introduced for the second and third tasks, it was not expected that times to first action would fall as

participants got accustomed to the system. This appears to be the case as mean times for how long participants took to make their first action generally did not decrease, and in the case of male participants were reasonably consistent (see Table 4).

	T1: Own Preference		T2: Box Limit		T3: Add Only	
	Male	Female	Male	Female	Male	Female
Mean	14.6	10.6	16.5	18.3	13.7	20.1
Std. Dev.	7.29	5.85	8.22	10.72	5.48	11.63

Table 4. Mean times (seconds) show that female participants were quicker to make their first action for the Own Preference task, but took longer for the Box Limit and Add Only tasks.

However, for female participants their mean times increased for each subsequent task. This resulted in mean times almost doubling between task 1 (10.6 seconds) and task 3 (20.1 seconds). The reason for which is evident in box plots for tasks 2 and 3 (see Figure 7), where high scores fall well outside the upper hinge.

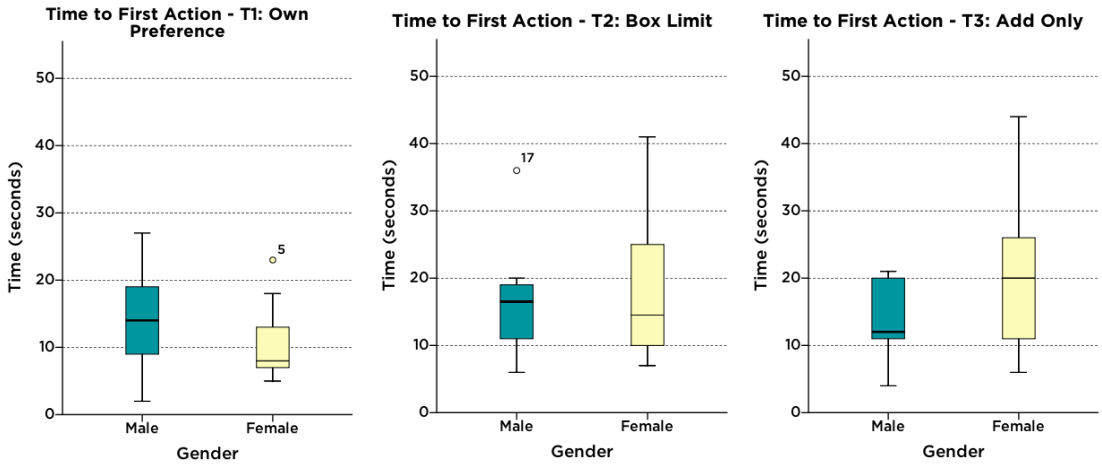


Figure 7. Box plots of Time to First Action, grouped by gender.

As interesting as these data may be, Mann-Whitney tests for each of the three tasks show these differences in time to first action to be statistically insignificant (see Table 5).

	T1: Own Preference	T2: Box Limit	T3: Add Only
Median (gender)	14 (m), 8 (f)	16.5 (m), 14.5 (f)	12 (m), 20 (f)
Mann-Whitney *U*	29.5	47.5	37.5

z	–1.556	–0.189	–0.952
p-value (exact sig. 2-tailed)	0.127	0.868	0.358

Table 5. Mann-Whitney tests found no statistical significance for each of the three tasks, for the time it takes males and females to make their first action to the recommender system.

4.5.2 First action choice

When using the recommender system, participants were able to perform two distinct action types, which were either exploratory or system changing (see Table 6).

<table>
<tr><th colspan="2">Exploratory Actions</th><th>System Changing Actions</th></tr>
<tr><th>Click</th><th>Scroll</th><th>Move</th></tr>
<tr><td rowspan="3">Clicking on a genre to view its contents</td><td rowspan="2">Scrolling programmes in All recommended row</td><td>Adding a programme to Like or Dislike box</td></tr>
<tr><td rowspan="2">Removing a programme from Like or Dislike box (Task 2 and 3 only)</td></tr>
<tr><td>Scrolling programmes in Like or Dislike box</td></tr>
<tr><td colspan="2"></td><td>Switching a programme from Like box to Dislike box or visa versa (Task 2 and 3 only)</td></tr>
</table>

Table 6. The first actions participants could perform on the system, which were either exploratory or system changing.

Participants therefore had the choice of whether to explore the recommender system to get an overview of the programmes available, or to jump straight in and start making changes to the system by adding, removing, or switching programmes. As each task was different this may have influenced the choice participants made.

For the Own Preference task (T1), both male and female participants favoured making a change to the system first over exploring it (see Figure 8). As this task was for their own personal preferences, they were able to dive straight in to make changes.

The Box Limit task (T2) saw the most extreme divergence between genders, as opposite choices were made. With female participants favouring exploration first, while male participants favoured system changing as their first choice (see Figure 9).

The Add Only task (T3), was almost the reverse of the Own Preference task (T1), with both genders this time favouring exploration. Though the degree that

exploration is favoured is not quite as strong as for change in task 1. It is not surprising that exploration was favoured for this task, due to the restriction that nothing could be removed from either the Like or Dislike box. Therefore, it was important to make the right choices of what to add that would counter what was already in either box.

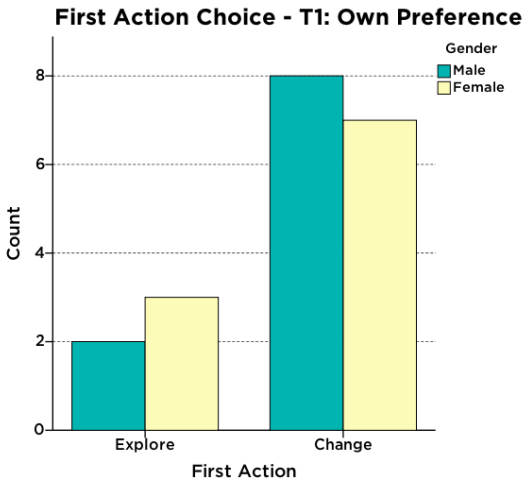


Figure 8. For the Own Preference task (T1), both male and female participants favoured making a change to the system as their first action.

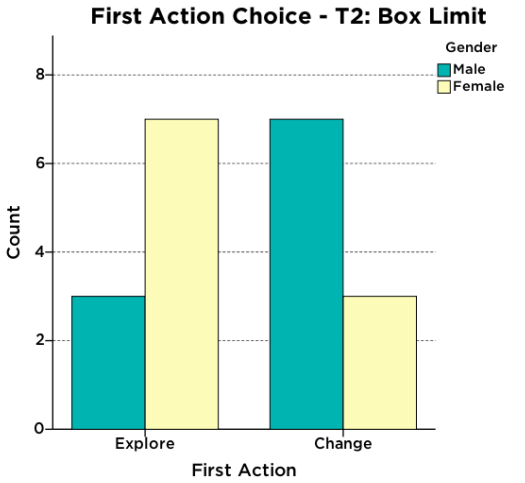


Figure 9. For the Box Limit task (T2), first actions by each gender were opposites.

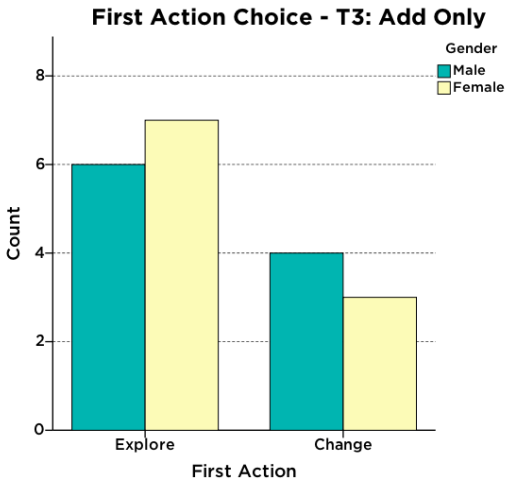


Figure 10. For the Add Only task (T3), both genders favoured system exploration as their first action.

Interestingly, for all tasks female participants outnumbered male participants in exploratory first actions, and male participants outnumbered female participants in system changing first actions. This aligns with theory (Meyers-Levy, 1989; Beckwith *et al.*, 2005b), which states females want to get an overview before making system changes, as well as being risk averse. And males take more risk so start making changes without an overview first.

From the Own Preference task (T1) data a 2 x 2 chi-square test was performed. However, since 50% of the cells had an expected frequency of less than 5, the appropriate statistical test was Fisher's Exact Probability. This gave *p-value* = 1 for a two-tailed hypothesis. The value of Cramer's V was 0.12, showing that the relationship between gender and first action choice was almost zero. The conclusion, therefore, is that there is no evidence to suggest an association between first action choice and gender.

From the Box Limit task (T2) data a 2 x 2 chi-square test was carried out to discover whether there was a significant relationship between gender and first action choice. The χ^2 value of 3.2 had an associated p-value = <0.074, DF = 1. Cramer's V was found to be 0.4 – thus 16% of the variation in frequencies of first action choice can be explained by gender. It can therefore be concluded that there is an insignificant association between first action choice and gender.

From the Add Only task (T3) data a 2 x 2 chi-square test was performed. However, since 50% of the cells had an expected frequency of less than 5, the appropriate statistical test was Fisher's Exact Probability. This gave *p-value* = 1 for a two-tailed hypothesis. The value of Cramer's V was 0.11, showing that the relationship between gender and first action choice was almost zero. The

conclusion, therefore, is that there is no evidence to suggest an association between first action choice and gender.

4.5.3 Time to first system change

For each of the three tasks, the time it took participants to make their first system change (FSC) was measured in seconds. In each of the tasks, female participants on average (mean) took longer to make a system change than male participants (see Figure 11–Figure 13).

The task which saw the closest mean times between genders (4.2 seconds) involved the participant setting the system up to recommend programmes that they would want to watch (T1: Own Preference). This differs from the other two tasks, as they both require the participant to set up the system to provide recommendations not for their own preferences, but for those of fictitious friends.

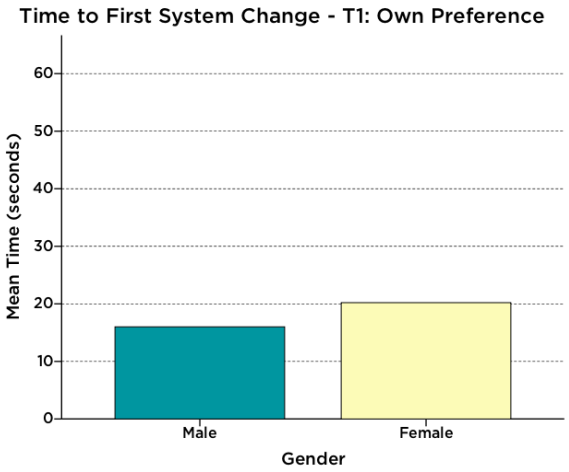


Figure 11. Female participants took on average (mean) 4.2 seconds longer than male participants to make their first system change on Task 1: Own Preference.

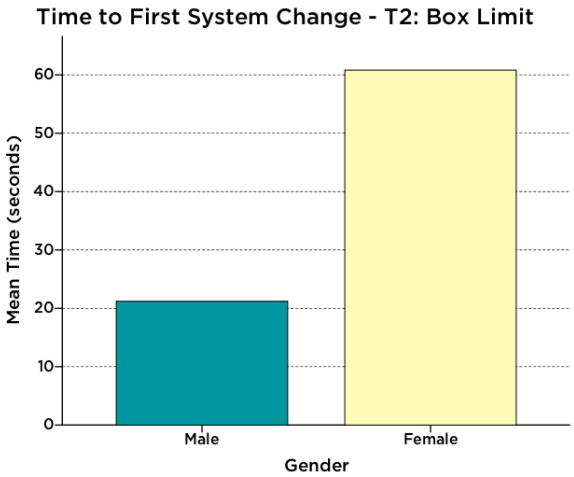


Figure 12. Female participants took on average (mean) 39.6 seconds longer than male participants to make their first system change on Task 2: Box Limit.

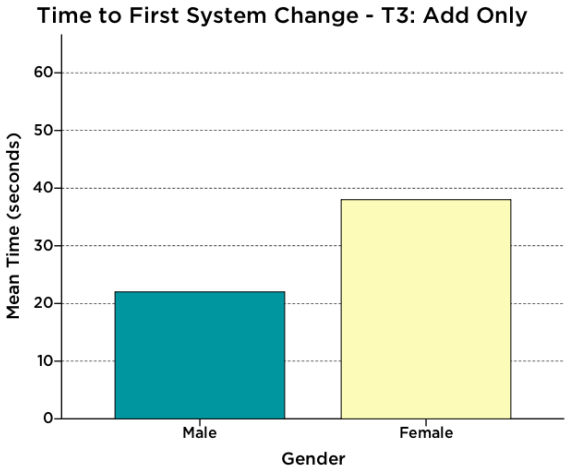


Figure 13. Female participants took on average (mean) 16 seconds longer than male participants to make their first system change on Task 3: Add Only.

	T1: Own Preference		T2: Box Limit		T3: Add Only	
	Male	Female	Male	Female	Male	Female
Mean Time (seconds)	16	20.2	21.2	60.8	22	38
Std. Dev.	8.29	17.06	15.53	32.02	12.75	30.53

Table 7. Time to make first system change across the three tasks—with females taking longer on average (mean) than males.

Histograms for time to first system change across the three tasks by both genders were inspected separately (see Figure 14). As data were skewed, and participant numbers were small, the most appropriate statistical test was the Mann-Whitney.

For the Own Preference task (T1), descriptive statistics showed (see Table 8) that female participants took less time (median = 12) to make a first change to the system, than male participants did (median = 15.5). The Mann-Whitney *U* was found to be 47.5 (z = –0.19) with an associated probability of 0.87 (exact sig. 2-tailed), suggesting the difference is not significant.

For the Box Limit task (T2), descriptive statistics showed that female participants took more time (median = 77) to make a first change to the system, than male participants did (median = 16.5). The Mann-Whitney *U* was found to be 17 (z = –2.5) with an associated probability of 0.01 (exact sig. 2-tailed), which

suggests significance. What does this actually mean then in terms of the interactions for that task. It is interesting that this was the task with the highest constraint of not going over three programmes in both the Like and Dislike boxes. Thus, was it down to risk and the fear—however nuanced—that they would be unable to achieve this, therefore holding back from making a system change. There is another possibility though, which is also backed up by theory, in that females want to get a more comprehensive understanding before they proceed (Meyers-Levy, 1989). Thus, it could be argued that the delay was due to wanting to ascertain more information first, either of how to proceed with a strategy, or to get a broader view of the programme choices that would be appropriate for that particular task.

Fortunately, however, we are not left to come up with theories of our own based purely on the timings obtained, as the participants watched back their use of the system with the intension of providing insight into their actions. The idea was to let them discuss what they were doing without much prompting (only when the participant went completely quiet—was a gentle prompt giving, asking them to explain what they were thinking at that point in the video). The following two participants were not needing of such prompts though—where this female participant had this to say about her delay in making a system changing decision for the box limit task (T2):

> "I think eventually - I think to start with I just kept looking at what was coming up and not sure what to do, and I think I was hesitating for quite a while before making any decisions of what to do." (P11)

Hesitation was also a factor for this next participant (P15) on the same task, who had this to say about her experienced, when she said:

> "I had to think carefully about my choices. I that's where I did hesitate - I did hesitate when I saw Legends - I know it is history, but I'm not quite - I think he would like it because it is to do with history and it's interesting."

Although, only two experiences, these do provide support for the risk-averse theory, with a restriction in place it did appear to the case that they wanted to take their time before making a move that could effect the choices then offered.

For the Add Only task (T3), descriptive statistics showed that female participants took more time (median = 24.5) to make a first change to the system, than male participants did (median = 16). The Mann-Whitney *U* was found to be 26.5 (z = –1.79) with an associated probability of 0.077 (exact sig.

2-tailed), which shows that it is possible that the slower times by females were not significant—thought quite close to be, again in relation to taking longer to make a first change to the system. This marginal result supports the previous statistically significant result.

Therefore, results from two of the Time to First System Change tasks lacked statistical significance by having probabilities greater than five percent. However, the significance of the Mann-Whitney test on the Box Limit task (T2) data provides enough evidence to suggest females take longer than males when making a first system change.

	T1: Own Preference	T2: Box Limit	T3: Add Only
Median (gender)	15.5 (m), 12 (f)	16.5 (m), 77 (f)	16 (m), 24.5 (f)
Mann-Whitney *U*	47.5	17	26.5
z	–0.19	–2.5	–1.79
p-value (exact sig. 2-tailed)	0.87	0.01	0.077

Table 8. Mann-Whitney test for Box Limit task, found high statistical significance for females taking longer than males when making a first system change.

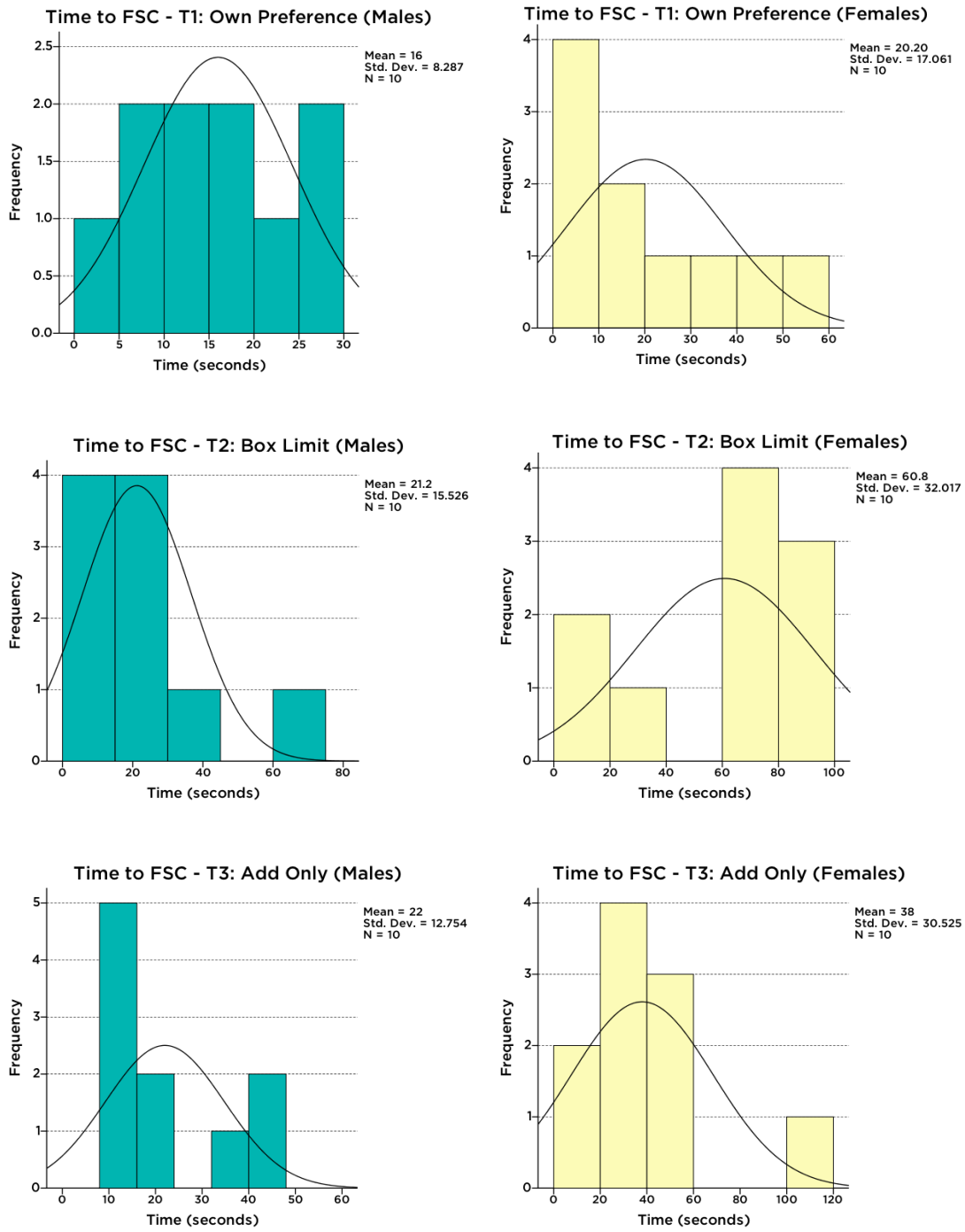


Figure 14. Histograms showing frequency of time (seconds) with which male and female participants took to make a first change to the system for each of the three tasks.

4.6 Results summary

It is wise to be cautious when reporting on findings from a study with a relatively low number of participants, where data sampling can skew the results. However, even with a sample of 20 participants, there are some results which deserve further thought. The fact that the self-efficacy scores were close

between males and females did suggest there might not be as many signs of gender differences due to many of the behaviours indicative of that construct. This however did prove to be unfounded, there were definite signs of behaviour patterns that suggest differences. Coming back to the self-efficacy scores, it seems unfair to question the data from a validated form that is a cornerstone of gender testing experiments.

Another aspect which is worth covering are the mental models of the participants. At the end of the sessions the participants were asked to explain how they thought the system providing the recommendations that it did—what it was doing "under the hood", so to speak. The results speak for themselves really, for example as this female participant explained:

> "I think that they probably have things put into genres and categories on their end. For example, if I moved everything from Sport into Dislike, I think they would be unlikely to show me other sports shows. And I would also think that it's not quite that broad, like the categories here like drama and entertainment, they would label them with more than just that, if that makes sense." (P3)

And this astute observation from another female participant, for which she was one of the only people to recognize the limitation of the system, and that the more you tell the system about your preferences, the fewer programmes that will remain which are similar:

> "I was doing well in the first task, and I had put a lot of shows that I liked or would like to watch in the Like, and a lot that I don't like in the Dislike - and then you asked me are these the five you would like to watch and I was like, no maybe only two of them, because everything I already like it in the Like and everything I dislike in the Dislike - so BBC need to come up with more shows to recommend me. Then I was like, I'm not so sure how this recommendation system would work, as the more information you feed it, it's supposed to become actuate. … That's the thing, it may be really actuate, but it runs out of things, so it's just - then the accuracy is decreasing."

Therefore, overall their understanding of the 'intelligence' of the system is fairly accurate, which could suggest that behavior deference's lie elsewhere than lacking confidence in their understanding of how to use it.

Chapter 5: Evaluation, reflections and conclusions

The overall aim of this research was to investigate whether gender differences exist in the use of recommender systems, and if so, what those differences are. This chapter will revisit each of the research objectives, give a summary of the findings, and provide self-reflection on the project process before finally concluding.

5.1 Evaluation of research objectives

1. **Identify current theory regarding known gender differences applicable to the use of intelligent interactive systems.**

The subject of gender differences in the field of human-computer interaction (HCI) covers abroad range of literature. This is due to HCI being a highly interdisciplinary area, covering amongst others computer science and psychology. For this reason, there are many different topics that could have been investigated that would have provided a higher-level of overview for the topic. However, with a limited amount of time for this research it was important to focus on the areas most appropriate to the specific objects—which were those specific to reported gender differences in software use, as well as use of intelligent interactive systems. The areas of current theory that seemed most relevant to this study, were self-efficacy, debugging, information processing theory, risk aversion, and tinkering.

2. **Evaluate critically the gender differences identified, including details on the tasks and interactions being performed.**

This objective also involved the literature review, though the focus was on comparing what was found across different studies and finding commonalities between the finding, as well as any differences that existed. Another aspect was to gain insight into the motivation behind actions that have been classed as specific gender differences. As it provided details on why the researchers assigned the action to one behavioural theory over another (e.g. risk aversion instead of information processing theory).

3. **Investigate how participants interact with a prototype recommender system, including their reasoning for specific actions.**

An experimental study was conducted successfully with 20 participants (10 males and 10 females), collecting both qualitative and quantitative data, using a between-participants design. Data included timings and frequencies observed from video-recordings of the participants use of the recommender system. Due to the nature of the study, which was designed to take a number of different measures, this meant that focus could not be given to a specific facet (e.g. risk aversion)—which may have proved beneficial, by building more evidence for just that one facet.

Retrospective think-aloud sessions were used to obtain qualitative data, providing insights into the motives behind the actions made by the participants. Unfortunately, due to time constraints more analysis of these data was not possible.

The data acquired through the self-efficacy questionnaire was considered unusual, in that male and female ratings were very similar. The reason for which could be an interesting direction for further research.

4. **Discover, through the analysis of data collected, whether gender differences exist, describing any that are found.**

Analysis of the empirical research data used statistical tests, these were mostly non-parametric (i.e. Mann-Whitney *U*-tests) due to the histograms showing that the data was not normally distributed. This was unfortunate due to more power being required to show significance, especially considering some of the findings were marginal.

Dealing with quantitative analysis for the first time was a challenging prospect. This was due to the necessity of working with statistics—grasping the important concepts, reading data, knowing which tests to perform and why. In addition, there was a new software program to learn, SPSS Statistics. However, it was quite satisfying to learn a new skill and gain an understanding, and appreciation, for quantitative analysis.

5.2 Self-reflection

5.2.1 Project timescale

Due to a challenging combination of events, it meant that the overall time frame for this project was shorter than usual. Which, as expected impacted what was achievable in the time remaining; requiring the whole process to remain very focused, with little or no room to explore areas that ‘might’ be of relevance.

Schedules were closely followed, working long hours seven days a week, in order to meet the objectives.

5.2.2 Participants

Recruitment of participants was harder than expected. This was not helped by the short window for recruitment and then to carry out the empirical data collection. When this project was first devised and proposed (see Appendix A), the aim was to recruit a total of 40 participants (20 males and 20 females). This however, due to time constraints and difficulties in attracting suitable participants, resulted in a reduced sample size, now consisting of 20 participants in total.

5.2.3 Planning

Having clearly defined procedures in place, documented and at hand, was highly beneficial. This was an aspect of the project that worked extremely well, with the session sequence being a prime example. Spending the time to consider in detail exactly how the sessions should run, and performing a pilot to check the running order works as it should and for the length of time expected. Having all necessary details for the sessions gathered in one document (Researcher Instructions) made things run smoothly and to the planned timescales.

5.2.4 Tasks and software

The recommender software used for the experiment, with hindsight, could have been improved in terms of its functionality and usability. This was most evident during the Box Limit task (T2), where participants were instructed to limit the amount of programmes in the Like and Dislike boxes to a maximum of three at any one time. This was difficult for four of the participants (P6, P11, P12, P15), of which two (P6, P11), were not successful in that task due to adding too many programmes. The other two participants made the odd slip, which they rectified immediately. It was not idea to make the participants remember this limitation while processing other information, because it places too much of demand on their working memory. Conversely, it would have been better if the system had been updated to not allow more than the permitted number of likes and dislikes.

Another change to the recommender system that would have been beneficial for the tasks, would have been to have had the row of recommended programmes for the All category, to be visible at all times, with a second row of programmes available to view different genres. That way, when participants made a system change while in a genre (e.g. added a programme to the Like

box), they would see immediately the effect that change had to the recommended programmes—rather than having to return to the All category to find out.

5.3 Conclusions

5.3.1 General conclusions

This research has identified several interesting findings, even if not statistically significant, such as data not reinforcing certain gender differences (self-efficacy), and that both males and females had strong mental models of the system. Although, what has been found does have validity, it would be interesting to perform the study again with a much large sample of participants. There were signs within the data, with marginal results, which could hold more power, with higher numbers.

The results of this research project have made a useful contribution to the gender HCI domain, providing further evidence as to a significant gender difference in the use of a recommender system.

5.3.2 Implications

To find statistical and psychological significance with delays in making system changes is worth of further exploration, in addition to marginal results in the same area, demonstrates a clear need for further research to take place. This is further compounded when considering the monumental number of people who use recommender systems every day—with Google alone receiving over 3.5 billion searches per day (Internetlivestats.com, 2015)—meaning the existence of gender differences in the use of these systems could potentially effect millions of people, thus making the need for further research into this domain worthwhile.

References

Arcand, M. and Nantel, J. (2012) 'Uncovering the Nature of Information Processing of Men and Women Online: The Comparison of Two Models Using the Think-Aloud Method', *J. theor. appl. electron. commer. res.*, 7(2), pp. 19-20.

Bandura, A. (1986) *Social foundations of thought and action: a social cognitive theory*, Prentice-Hall, Englewood Cliffs, N.J.

Bbc.co.uk, (2015) *BBC - Research and Development: Sibyl Recommender Update*. Available at: http://www.bbc.co.uk/blogs/researchanddevelopment/2012/11/sibyl-recommender-update.shtml (Accessed: 2 August 2015).

Beckwith, L., Burnett, M., Grigoreanu, V., and Wiedenbeck, S. (2006a) 'Gender HCI: What about the software?'. *IEEE Computer 39, 11,* pp. 83-87.

Beckwith, L., Burnett, M., Wiedenbeck, S., Cook, C., Sorte, S. and Hastings, M. (2005a) 'Effectiveness of end-user debugging features: Are there gender issues?' *In Proc. of the ACM Conference on Human Factors in Computing Systems*, pp. 869–878.

Beckwith, L., Inman, D., Rector, K. and Burnett, M. (2007) 'On to the real world: Gender and self-efficacy in Excel'. *Proc. VL/HCC 2008,* IEEE, pp. 119-126.

Beckwith, L., Kissinger, C., Burnett, M., Wiedenbeck, S., Lawrance, J., Blackwell, A. and Cook, C. (2006b) 'Tinkering and gender in end-user programmers' debugging'. *ACM Conference on Human Factors in Computing Systems*, pp. 231-240.

Beckwith, L., Sorte S., Burnett, M., Wiedenbeck S., Chintakovid T., and Cook C. (2005b) 'Designing features for both genders in end-user programming environments'. *Proc. VL/HCC,* IEEE, pp. 153-160.

Bunt, A., Lount, M. and Lauzon, C. (2012) 'Are explanations always important?', *Proc. IUI*, pp. 169–178.

Busch, T. (1995) 'Gender differences in self-efficacy and attitudes toward computers'. *Journal of Educational Computing Research 12*, 2, pp. 147-158.

Compeau, D. and Higgins, C. (1995) 'Application of social cognitive theory to training for computer skills', *Information Systems Research*, *6,2*, pp. 118-143.

Croson, R. and Gneezy, U (2009) 'Gender Differences in Preferences.' *Journal of Economic Literature 47(2)*, pp. 448-474.

Dancey, C. and Reidy, J. (2014) *Statistics Without Maths for Psychology*, *(6th ed.).* Pearson Education, Limited, Essex, UK.

Encyclopedia Britannica, (2015) *construct | psychology*. Available at: http://www.britannica.com/science/construct (Accessed: 12 August 2015).

Grigoreanu, V., Cao, J., Kulesza, T., Bogart, C., Rector, K., Burnett, M. and Wiedenbeck, S. (2008) 'Can feature design reduce the gender gap in end-user software development environments?' *Proc. VL/HCC,* IEEE, pp. 149-156.

Hart, S. and Staveland, L. (1998) 'Development of a NASA-TLX (Task load index): Results of empirical and theoretical research', Hancock, P. and Meshkati, N. (Eds.), *Human Mental Workload*, pp. 139-183.

Internetlivestats.com, (2015) *Google Search Statistics - Internet Live Stats*. Available at: http://www.internetlivestats.com/google-search-statistics/ (Accessed: 17 September 2015).

Kahneman, D. and Tversky, A. (2000) *Choices, values, and frames*. New York: Russell sage Foundation.

Kulesza, T., Burnett, M., Wong, W.-K. and Stumpf, S. (2015) 'Principles of Explanatory Debugging to Personalize Interactive Machine Learning'. *In Proc. IUI*, pp. 126–137.

Kulesza, T., Stumpf, S., Burnett, M., and Kwan, I. (2012) 'Tell me more? The effects of mental model soundness on personalizing an intelligent agent'. *In Proc. HF/CS*, pp. 1–10.

Kulesza, T., Stumpf, S., Burnett, M., Wong, W.-K., Riche, Y., Moore, T., Oberst, I., Shinsel, A. and McIntosh, K. (2010) 'Explanatory debugging: Supporting end-user debugging of machine-learned programs'. In *Proc. VL/HCC*, IEEE, pp. 41-48.

Kulesza, T., Wong, W.-K., Stumpf, S., Perona, S., White, R., Burnett, M. M. and Ko, A. J. (2009) 'Fixing the program my computer learned: Barriers for end users, challenges for the machine'. *In Proc. IUI*, pp. 187–196.

Melville, P., Sindhwani, V. (2010) *Recommender Systems, Encyclopedia of Machine Learning*. Berlin: Springer.

Meyers-Levy, J. (1989) Gender differences in information processing: A selectivity interpretation. In P. Cafferata & A. Tybout, (Eds) *Cognitive and Affective Responses to Advertising.* Lexington, Ma, Lexington Books.

Nelson, J. (2012) 'Are Women Really More Risk-Averse than Men?', *SSRN Journal.* doi: 10.2139/ssrn.2158950.

Oates, B. (2006) *Researching information systems and computing*. London: SAGE Publications.

Oxford Dictionaries, (2015) *gender - definition of gender in English from the Oxford dictionary*. Available at: http://www.oxforddictionaries.com/definition/english/gender (Accessed: 12 September 2015).

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. (2011) *Recommender systems handbook*. New York: Springer.

Rowe, M. (1978) *Teaching Science as Continuous Inquiry: A Basic (2nd ed.).* McGraw-Hill, New York, NY.

Sibyl.prototyping.bbc.co.uk, (2015) *Sibyl Experimental Recommender*. Available at: http://sibyl.prototyping.bbc.co.uk (Accessed: 2 August 2015).

Wikipedia, (2015) *List of most popular websites*. Available at: https://en.wikipedia.org/wiki/List_of_most_popular_websites (Accessed: 14 September 2015).

World Health Organization, (2015) *WHO | Gender*. Available at: http://www.who.int/mediacentre/factsheets/fs403/en/ (Accessed: 17 September 2015).

Appendix A: Proposal

Name:	**Paul Galbraith**
Email Address:	**paulgalbraith@gmail.com**
Contact Phone Number:	**(+44) 07740 353 466**
Project Title:	**Gender differences in using low-cost intelligent interactive systems**
Supervisor:	**Simone Stumpf**

Introduction

There are many types of low-cost intelligent interactive systems (IIS), such as Facebook friend finder, Google Suggest (see figure 1), IMDb movie finder, YouTube recommended videos (see figure 2), Amazon and iTunes Genius. It is not clear when using these what is going on behind the scenes to deliver the results back that they do.

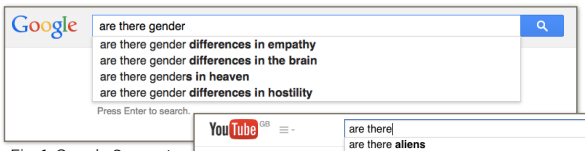


Fig. 1: Google Suggest

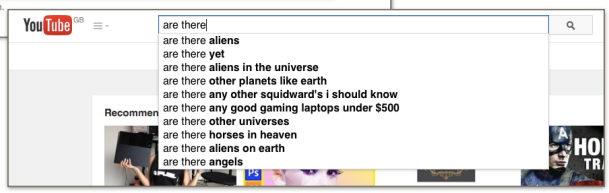


Fig. 2: YouTube recommended videos

Due to the uncertainty in how these systems work, interacting with them could lead to certain thoughts by the user, such as how useful these are, how are they getting the results, could I change my behaviour to improve the accuracy of the results returned.

However, it gets more complicated, theory says that we should expect some differences between how different genders use intelligent systems, one example being that women have been found to be more risk averse than men (Croson, R. and Gneezy, U, 2009). This could therefore result in differences with how these systems are perceived, understood and used by each gender.

Project Objectives

In order to investigate whether there are differences between genders when using IIS, two main research questions will be asked:

- Are there gender differences in using low-cost intelligent interactive systems?
- And if so, what are those differences?

To answer these questions, a series of additional questions will need to be answered that clarify how these systems are used, such as: what the

perceived utility is when using these types of recommenders; what is the mental model of the users, and do they align with how the system actually functions; if the users have a desire for additional information about how the system works; and if the users have a desire to "correct" how the system works, so that it may be more suited to their needs.

The data would then be used to look for patterns and identity certain behaviours in relation to the various questions asked. These could then be grouped by gender to see if answers of certain types were more frequent by a certain gender.

The result of which will be a better understanding of how IIS are used by both genders, and if there are patterns of use between the two. These could include insight into whether these systems need to be changed to work better for either or both genders.

Why and For Whom Will it be Useful

As this research sets out to explore whether there are gender differences when using IIS, the beneficiaries would include software developers—for if there are differences found—these could have implications for the design of other systems that involve some kind of uncertainty. They could then make changes to their applications to meet the needs of both genders, which could impact use and thus sales (for commercial software).

Obviously, this would also impact the users of these systems, both in terms of their understanding of the differences between users (if any are found), and if systems are changed as a result, these may allow for the types of interactions suited to both genders, rather than just one.

Critical Context

Confidence and self-efficacy have been explored in terms on gender differences, though this was with bugging of spreadsheets (Beckwith *et al.*, 2005). This research did find that there were differences between genders, finding that "Females had lower self-efficacy than males did about their abilities to debug." and "Females were less likely than males were to accept the new debugging features." (Beckwith *et al.*, 2005).

Suggestions are made that designers of software products should make appropriate accommodations for gender differences, or else if could affect how willing a group of users are to accepting new features that could benefit them (Beckwith *et al.*, 2005).

Another study that looked at how users attempted to fix an email tool that files emails according to certain criteria. Here they found that "Gender differences were present in the number of barriers encountered, the sequence of barriers, and usage of debugging features." (Kulesza et al., 2009).

It should be mentioned, that although research exists that does find gender differences, for example female students having lower self-efficacy than male students when completing complex tasks within word processing and

spreadsheets software (Busch, 1995), generalised statements should be avoided.

This is discussed by Nelson (2012), who suggests "The statement 'women are more risk averse than men' is fundamentally a metaphysical assertion about unobservable essences or characteristics, and therefore cannot be empirically proven or disproven.". She then goes on to say "The widespread acceptance of such statements appears to perhaps be rooted more in confirmation bias than in reality." (Nelson, 2012).

In terms of low-cost recommenders which are the type of IIS that are the focus of this project, earlier research has found that "none of the applications had any explanation facilities to speak of." (Bunt, Lount and Lauzon, 2012). They also looked for "possible triggers for participants' desire for more information", and found reasons for such triggers include understanding inconsistent, or good behaviour, wanting to improve the interactions, or just because they were curious.

For these reasons conducting research into a low-cost intelligent interactive systems, in terms of gender differences—would be a worthwhile piece of research as if there are gender differences when using a low-cost IIS, as opposed to something that one would expect risk and thus self-efficacy to have more impact. Then it will have implications for the design of other systems that involve some kind of uncertainty.

Further literature research

It is understood that the above is only a initial search across the available texts, therefore, a wider literature review will be performed to expand on the areas touched upon above, including those of gender differences in using a variety of systems, before concentrating on the specific system type which is the focus of this proposed project—intelligent interactive systems. Research pertaining to whether users find these systems useful, the mental models users have of how they operate, if additional information is wanted by users, and also if users would want to correct these types of systems will all be sought.

These texts will be located by compiling a list of key search terms to use across a range of databases, while maintaining a record of the terms searched and against which databases. References and how often a paper has been cited will be used to consider appropriateness of material. Relevant texts may also be recommended by the project supervisor, as well as other academics.

Approaches: Methods & Tools for Evaluation & Analysis

DATA COLLECTION

In order to answer the research questions, detailed behaviour will need to be captured of how users interact with IIS. One means of achieving this would be to have participants self-report their interactions with IIS (Bunt, Lount and Lauzon, 2012). This would have the benefit of capturing data about users' real-world use of these systems. However, with this approach there is the danger of the participants choosing which details to include and exclude

from their reports. Also, the participants may not even be aware of, or be able to recall, all of their actions and thoughts—especially if reporting occurs later that day.

An alternative approach is to do an observational study, where participants are asked to think-aloud while performing tasks using deployed, IIS. This has the following benefits:

- A greater level of consistency as all users will be performing the same tasks, and using the same IIS
- Everything the participants do and say is observed by the researcher, and recorded so that it be analysed in detail afterwards

Online Questionnaire

In order to set appropriate tasks that match real-world uses of IIS an online questionnaire will be used to gather this information. Participants will be first asked the following question:

- Which of the following recommender systems have you used in the past week? (The participant will be able to select as many as they want from a list of IIS, as well as allowing them to add their own if not included on the list)

For each IIS the participant selects, they will be asked to complete questions about it, including:

- What did you last use [name of recommender system] for? (if they can't remember, they will be asked to recall any use from the past, or what they may expect to use if for in the future)
- Can you recall a time you used the system and were unsatisfied with the recommendations it gave? Please give details (this is needed so there can be a mix of good and bad recommendations by the IIS, to see how participants respond to both)

The last part of the questionnaire will ask participants for their age-range, sex and country. These will assist with task selection if patterns are found (e.g. if certain types of tasks are favoured by women, or men, a balanced mix of each can be used during the observations).

A pilot of the questionnaire will be conducted first to ensure all the questions are understood and that the answers given are suitable for the purpose of creating real-world tasks.

A minimum of 50 completed questionnaires will be sought to get a range of responses, though it will stay active for one week even if the 50 are reached sooner to get as many responses as possible.

Observational Study

A pilot of the observational study will be conducted first to ensure all the questions and tasks are understood, and that the data gathered is suitable for the purpose of answering the research questions.

For the observations, 40 participants (20 female, 20 male) will be recruited. Observing 20 users of each gender will allow enough qualitative data to be generated to make formative comparisons between the two groups.

The participants will be found through contacts, social media and university notice boards, and if necessary, will be offered compensation for their participation. Screening will be used to find suitable participants by excluding those that:

- Use fewer than two IIS or are infrequent users
- May have trouble committing to the study (e.g. unsure if they will be available during the timeframe the observations will take place, not wanting to give up an hour of there time, plus travel time)
- Could have an understanding of how these systems work, those with a computer science background will be excluded from participating in the study

The participants will all use a supplied laptop for the study, rather than their own desktop, laptop, tablet or mobile. This is for practical purposes to have software installed to record the screen and user while they perform the tasks. And seeing that the tasks will only involve users inputting text into a search field for the IIS to give them recommendations, then there is not likely to be much impact from them not using their own device in terms of familiarity.

The browser's cache will also be emptied after each participant and cookies deleted. This is to ensure that the IIS will not have retained any data from a previous user that may influence what it recommends. This will of course mean that the IIS may behave differently in what it recommends compared to on the user's own device (as the IIS may retain data about the user's previous searches to improve the current search). This however will ensure all users are using the IIS from the same starting point, ensuring consistency between users for when comparisons are made.

Prior to the users completing the tasks, they will be asked a series of questions to to gain an overview of their experience with IIS, including:

- Which recommender systems they have used in the past month
- How often they use these recommender systems
- What devices they normally use these recommender systems on

The actual observations will consist of several tasks, the nature of these will depend on the data from the questionnaires. However, the idea is to cover several different IIS and to have at least three tasks per IIS. The actual time to type queries into the search fields and to discuss the recommendations the IIS give will likely be fairly quick—though the pilot will confirm this and allow for task numbers to be set to ensure the participants are not required for longer than an hour.

The users will be told a scenario and given a task to perform. These will be created using the questionnaire results data, but could be similar to the following:

- Scenario: In the evenings you often spend 20 mins on YouTube watching videos that make you laugh
- Task: Use YouTube's recommender system to find an appropriate video

The users will be asked to think-aloud as they perform the tasks, to describe what they are doing and why, as well as anything else they are thinking. So apart from asking them to do that, they will not be prompted for additional types of information (e.g. how do you think the system works). For the purpose of the study, it is important that the users' comments about the IIS are their own. Therefore, the facilitator will only prompt users if they go quiet, by asking "what are you thinking now?".

What the users say and do will be captured by software on the laptop, to be analysed later. This will allow the facilitator to focus on the user and task at hand.

EVALUATION & ANALYSIS

Each think-aloud during the observations will be transcribed in full, with transcripts analysed using qualitative analysis techniques. This will be achieved by "abstracting from the research data the verbal, visual or aural themes and patterns" (Oates, 2006, chapter 18). However, three common themes for IIS have already been found, and will be used alongside a proposed fourth—these are: perceived utility, nature of mental models, desire for additional information, and desire to adapt system behaviour (Bunt, Lount and Lauzon, 2012).

The themes identified from the observation data will be used quantitatively, to allow comparisons to be made between gender groups. For example, how many users desired additional information and what was the split between men and women.

However, as the dataset is small it will be made clear that these should not be used to infer any findings are representative of more than just the participants involved in the study. Though if patterns are found, they could warrant further investigation.

All data will be analysed for each individual participant first, and only later will the results be compared by gender to discover any differences between the two groups. This will be done to reduce any unintentional bias that may come from analysing participants already grouped according to gender.

Risks & Limitations

If the plan of work that follows is adhered to, then the project can be completed in the proposed timeframe—as the duration of the various tasks are considered achievable. Thus, bar a serious accident happening to the project researcher or another force majeure event, the 25th September 2015 deadline can be meet.

All interviews and any other meetings with participants will take place in public spaces (e.g. university common rooms, library study rooms, cafes) for safety purposes and to provide a relaxed environment.

RISK	LIKELIHOOD	IMPACT	MITIGATION
Unable to find participants	Low	High	Will start recruiting participants early using various avenues (e.g. contacts, social media, university notice boards), and will increase compensation as necessary
Not achieving balanced gender groups	Medium	High	Recruiting early should alleviate this, as it will allow time to find additional participants of a specific gender group if required
Participants withdraw prior to study	Medium	Medium	Build a pool of reserve participants to have on standby if needed
Participants withdraw during study	Medium	High	Will screen potential participants to ensure suitability, and the number of participants chosen will be higher than necessary to allow for potential drop-outs
Lack of diverse intelligent interactive systems	Low	Low	Decide at time whether to accept the low diversity of IIS or use additional participants from the reserve pool
Computer failure or loss	Low	High	All data will be encrypted and backed up to an external drive frequently

Ethical Issues

In terms of ethical considerations, all participants will be over the age of 18, not classed as vulnerable, and will be fully informed and giving of consent.

A completed ethics checklist can be found at the end of this document, which contains full details of all ethical issues considered in the scope of this proposed research project. In addition, a sample participant information sheet and a sample consent form are included.

Work Plan

PROJECT TASK DESCRIPTION	29/06/15	06/07/15	13/07/15	20/07/15	27/07/15	03/08/15	10/08/15	17/08/15	24/08/15	31/08/15	07/09/15	14/09/15	21/09/15
Review Literature	■	▲											
Recruitment process (for pilots and actual study)	■	■	■										
Prepare questionnaire		■											
Conduct questionnaire pilot		■											
Analyse questionnaire data and adjust questions if necessary			■										
Prepare observation material			■										
Conduct observation pilot				■									
Analyse observation data and adjust tasks/format if necessary				■									
Conduct observation study					■	▲							
Transcribe think-alouds in full						■	■						
Analyse observation data								■	■				
Finish writing dissertation										■	▲		
Proofread dissertation												■	
Submit project report 25/09/15													▲

▲ = milestone

The plan of work is laid out over weekly blocks (Monday-Sunday), with the first week starting Monday 29th June 2015 (project will officially begin on 1st July of that week) and the final week starting Monday 21st September 2015 (deadline for project report Friday 25th September 2015).

Supervisor meetings will be held every two weeks (time and dates to be agreed), for the three month duration. Writing of the dissertation will be ongoing throughout the three month period, with sections drafted as and when possible following the completion of project tasks—allowing the project supervisor to read and provide feedback at regular intervals.

References

Beckwith, L., Burnett, M., Wiedenbeck, S., Cook, C., Sorte, S. and Hastings, M. (2005) 'Effectiveness of end-user debugging features: Are there gender issues?' *In Proceedings of the ACM Conference on Human Factors in Computing Systems*. 869–878.

Bunt, A., Lount, M. and Lauzon, C. (2012) 'Are explanations always important?', *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces - IUI '12*, pp. 169–178.

Busch, T. (1995) 'Gender differences in self-efficacy and attitudes toward computers'. *Journal of Educational Computing Research 12*, 2, pp. 147-158.

Croson, R. and Gneezy, U (2009) 'Gender Differences in Preferences.' *Journal of Economic Literature 47(2)*, pp. 448-474.

Kulesza, T., Wong, W., Stumpf, S., Perona, S., White, R., Burnett, M., Oberst, I., Ko, A. (2009) 'Fixing the program my computer learned: Barriers for end users, challenges for the machine'. *In Proc. IUI*, pp. 187–196.

Nelson, J. (2012) 'Are Women Really More Risk-Averse than Men?', *SSRN Journal.* doi: 10.2139/ssrn.2158950.

Oates, B. (2006) *Researching information systems and computing*. London: SAGE Publications.

Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with ethical guidelines. In some cases ethics approval will have to be obtained from an ethics committee before the project can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that due consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

Part A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

Part B: Ethics Proportionate Review Form. This part is an application for ethical approval of low-risk research. Students who have answered "no" to questions 1 – 18 and "yes" to question 19 in the checklist must complete this part. The project supervisor has delegated authority to approve this application.

Part A: Ethics Checklist

If your answer to any of the following questions (1 – 3) is YES, you must apply to an appropriate external ethics committee for approval:		*Delete as appropriate*
1.	Does your project require approval from the National Research Ethics Service (NRES)? (E.g. because you are recruiting current NHS patients or staff? If you are unsure, please check at http://www.hra.nhs.uk/research-community/before-you-apply/determine-which-review-body-approvals-are-required/)	**No**
2.	Will you recruit any participants who fall under the auspices of the Mental Capacity Act? (Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee http://www.scie.org.uk/research/ethics-committee/)	**No**
3.	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? (Such research needs to be authorised by the ethics approval system of the National Offender Management Service.)	**No**

If your answer to any of the following questions (4 – 11) is YES, you must apply to the Senate Research Ethics Committee for approval (unless you are applying to an external ethics committee):		*Delete as appropriate*
4.	Does your project involve participants who are unable to give informed consent, for example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf?	**No**
5.	Is there a risk that your project might lead to disclosures from participants concerning their involvement in illegal activities?	**No**
6.	Is there a risk that obscene and or illegal material may need to be accessed for your project (including online content and other material)?	**No**
7.	Does your project involve participants disclosing information about sensitive subjects?	**No**
8.	Does your project involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning? (http://www.fco.gov.uk/en/)	**No**
9.	Does your project involve invasive or intrusive procedures? For example, these	**No**

	may include, but are not limited to, electrical stimulation, heat, cold or bruising.	
10.	Does your project involve animals?	**No**
11.	Does your project involve the administration of drugs, placebos or other substances to study participants?	**No**

If your answer to any of the following questions (12 – 18) is YES, you must submit a full application to the Computer Science Research Ethics Committee (CSREC) for approval (unless you are applying to an external ethics committee or the Senate Research Ethics Committee). Your application may be referred to the Senate Research Ethics Committee.		*Delete as appropriate*
12.	Does your project involve participants who are under the age of 18?	**No**
13.	Does your project involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.	**No**
14.	Does your project involve participants who are recruited because they are staff or students of City University London? For example, students studying on a particular course or module. (If yes, approval is also required from the Head of Department or Programme Director.)	**No**
15.	Does your project involve intentional deception of participants?	**No**
16.	Does your project involve participants taking part without their informed consent?	**No**
17.	Does your project pose a risk to participants or other individuals greater than that in normal working life?	**No**
18.	Does your project pose a risk to you, the researcher, greater than that in normal working life?	**No**

If your answer to the following question (19) is YES and your answer to all questions 1 – 18 is NO, you must complete part B of this form.		
19.	Does your project involve human participants? For example, as interviewees, respondents to a questionnaire or participants in evaluation or testing.	**Yes**

Part B: Ethics Proportionate Review Form

If you answered YES to question 19 and NO to all questions 1 – 18, you may use this part of the form to submit an application for a proportionate ethics review of your project.

The following questions (20 – 24) must be answered fully.		*Delete as appropriate*
20.	Will you ensure that participants taking part in your project are fully informed about the purpose of the research?	**Yes**
21.	Will you ensure that participants taking part in your project are fully informed about the procedures affecting them or affecting any information collected about them, including information about how the data will be used, to whom it will be disclosed, and how long it will be kept?	**Yes**
22.	When people agree to participate in your project, will it be made clear to them that they may withdraw (i.e. not participate) at any time without any penalty?	**Yes**
23.	Will consent be obtained from the participants in your project? Consent from participants will be necessary if you plan to gather personal data. "Personal data" means data relating to an identifiable living person, e.g. data you collect using questionnaires, observations, interviews, computer logs. The person might be identifiable if you record their name, username, student id, DNA, fingerprint, etc. *If YES, attach the participant information sheet(s) and consent request form(s) that you will use. You must retain these for subsequent inspection. Failure to provide the filled consent request forms will automatically result in withdrawal of any earlier ethical approval of your project.*	**Yes**
24.	Have you made arrangements to ensure that material and/or private information obtained from or about the participating individuals will remain confidential? Provide details: Names and other identifiable details of participants will be kept separate from other data collected where possible, using I.D. numbers to associate the two. Paper documents will be stored on secure private premises, in a locked container. Digital files will be encrypted (where possible) and stored on a password-protected personal computer and/or encrypted external hard drive stored on secure private premises.	**Yes**

If the answer to the following question (25) is YES, you must provide details		*Delete as appropriate*
25.	Will the research be conducted in the participant's home or other non-University location? *If **YES**, provide details of how your safety will be ensured:* Any interviews or meetings conducted outside of university premises will be in public spaces, such as cafes. This is to ensure there are other people present for safety reasons.	**Yes**

Attachments (these must be provided if applicable):	*Delete as appropriate*
Participant information sheet(s)	**Yes**
Consent form(s)	**Yes**
Questionnaire(s)**	**Not available**
Topic guide(s) for interviews and focus groups**	**Not available**
Permission from external organisations (e.g. for recruitment of participants)**	**Not applicable**

**If these items are not available or not applicable at the time of submitting your project proposal, preliminary approval through proportionate review can still be given. This will be subject to you submitting the items to your supervisor for approval at a later date. Approval must be obtained prior to the research commencing.

Templates

The University provides templates which should be used as the basis for your participant information sheets and consent forms. These are available from the links below but **must** be adapted according to the needs of your project before they are submitted for consideration.

Adult information sheet:

http://www.city.ac.uk/__data/assets/word_doc/0018/153441/TEMPLATE-FOR-PARTICIAPNT-INFORMATION-SHEET.doc

Adult consent form:

http://www.city.ac.uk/__data/assets/word_doc/0004/153418/TEMPLATE-FOR-CONSENT-FORM.doc

CITY UNIVERSITY LONDON

CONSENT FORM

Title of Study: Gender differences in using low-cost intelligent interactive systems

Please initial box

1.	I agree to take part in the above City University London research project. I have had the project explained to me, and I have read the participant information sheet, which I may keep for my records. I understand this will involve: • being interviewed and observed by the researcher • allowing the interview and observation to be videotaped/audiotaped	
2.	This information will be held and processed for the following purpose(s): To allow the researcher named below to complete his INM363 Individual Project I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party. No identifiable personal data will be published. The identifiable data will not be shared with any other organisation.	
3.	I understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdraw at any stage of the project without being penalized or disadvantaged in any way.	
4.	I agree to City University London recording and processing this information about me. I understand that this information will be used only for the purpose(s) set out in this statement and my consent is conditional on the University complying with its duties and obligations under the Data Protection Act 1998.	
5.	I agree to take part in the above study.	

_____	_____	_____
Name of Participant	Signature	Date
_____	_____	_____
Name of Researcher	Signature	Date

When completed, 1 copy for participant; 1 copy for researcher file.

CITY UNIVERSITY LONDON

City University London
Northampton Square
London EC1V 0HB
T: +44 (0)20 7040 www.city.ac.uk

CITY UNIVERSITY LONDON

PARTICIPANT INFORMATION SHEET

Gender differences in using low-cost intelligent interactive systems

I would like to invite you to take part in a research study. Before you decide whether you would like to take part it is important that you understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information.

What is the purpose of the study?
This study is part of a MSc in Human-Centred Systems course and is for my INM363 Individual Project. The project will run for three months from July 2015 to September 2015. It will investigate gender differences in using low-cost intelligent interactive systems.

Why have I been invited?
You have indicated that you use these types of intelligent interactive systems, and so would be suitable for this study.

Do I have to take part?
Participation in the project is voluntary, and you can choose not to participate in part or all of the project. You can withdraw at any stage of the project without being penalised or disadvantaged in any way.

It is up to you to decide whether or not to take part. If you do decide to take part you will be asked to sign a consent form. If you decide to take part you are still free to withdraw at any time and without giving a reason.

What will happen if I take part?

- There will be an observation study, including interview (lasting up to 1 hour)
- The research study will last for 3 months
- You will need to meet the researcher once
- Semi-structured questions will be used before and after the observation study. The observation will involve completing set tasks, while performing a think-aloud
- Data will be collected from the study (audio and video)
- The observation study will take place either at City University London, or a mutually agreeable public space (such as a café)

Expenses and Payments

- £10 Amazon voucher as compensation for your time

What do I have to do?
You will meet the researcher once, answering questions about your use of intelligent interactive systems. You will then be asked to perform tasks while performing a think-aloud, allowing the researcher to observe your interactions with intelligent interactive systems. There will be follow-up questions afterwards in relation to the tasks.

What are the possible disadvantages and risks of taking part?
No particular risks have been identified, but please let me know if you feel uncomfortable at any time or wish to stop the study.

What are the possible benefits of taking part?
Although there are no direct benefits to you taking part, I hope you will find participating interesting.

What will happen when the research study stops?
Findings from the study will be written up as INM363 Individual Project. Data will be stored securely, and your identifiable information will be kept separate from the information you provided.

Will my taking part in the study be kept confidential?

- Data used within the report will be anonymised and stored securely
- Only the researcher, and if requested, university bodies will have full access to the data and participant details
- Data will be stored in a secure location (digital and paper) and will be destroyed when no longer required by the university

What will happen to results of the research study?
Findings from this study will be used at INM363 Individual Project, and used for university assessment, it will not be possible for you to obtain a copy (although you can contact me by e-mail at paul.galbraith@city.ac.uk if you would like me to send you a summary of my findings related to your interview and diary study).

What will happen if I don't want to carry on with the study?
You are free to withdraw from the study without an explanation or penalty at any time.

What if there is a problem?
If you have any problems, concerns or questions about this study, you should ask to speak to a member of the research team. If you remain unhappy and wish to complain formally, you can do this through the University complaints procedure. To complain about the study, you need to phone 020 7040 3040. You can then ask to speak to the Secretary to Senate Research Ethics Committee and inform them that the name of the project is: Gender differences in using low-cost intelligent interactive systems

You could also write to the Secretary at:
Anna Ramberg
Secretary to Senate Research Ethics Committee
Research Office, E214
City University London
Northampton Square
London
EC1V 0HB
Email: Anna.Ramberg.1@city.ac.uk

City University London holds insurance policies which apply to this study. If you feel you have been harmed or injured by taking part in this study you may be eligible to claim compensation. This does not affect your legal rights to seek compensation. If you are harmed due to someone's negligence, then you may have grounds for legal action.

Who has reviewed the study?
This study has been approved by City University Research Ethics Committee.

Further information and contact details
For more information about this study, or if you have any queries or concerns, contact Dr Simone Stumpf (Simone.Stumpf.1@city.ac.uk).

Thank you for taking the time to read this information sheet.

CITY UNIVERSITY LONDON

City University London
Northampton Square
London EC1V 0HB
T: +44 (0)20 7040 www.city.ac.uk

Appendix B: Email recruitment advert

Hello,

My name is Paul Galbraith and I am currently looking for a number of people to participate in a research study for my MSc in Human-Centred Systems at City University London. The study concerns how users interact with recommender systems.

Who can take part?
I am looking for participants who watch several hours of TV programmes a week, particularly those familiar with BBC programmes. You should also be a regular user of a computer, tablet, or smartphone – visiting a variety of websites every week (e.g. YouTube, Facebook, Amazon, eBay). Unfortunately, you are unable to take part if you have a background in recommender system algorithms/design. And please feel free to forward this email onto anyone you know that may be interested in participating.

What will happen during the study?
You will be asked to participate in an observational study (including a retrospective think-aloud), complete two short questionnaires, and a semi-structured interview. During this time, a number of questions will be asked to understand your use of an interactive recommender system. For the observation you will be asked to perform tasks on a supplied laptop, that will involve using an interactive recommender system. The laptop's screen will be recorded using screen capture software, along with an audio-recording of what you say during the session.

Where and when will the study take place?
This research will take place at City University London, and will start week commencing Monday 17th August. You will only be required to attend one session at a timeslot agreed upon by the researcher and yourself. The research session will last up to an hour.

Will your taking part in the study be kept confidential?
Data used within the report will be anonymised and stored securely. Only the researcher, and if requested, university bodies will have full access to the data and participant details. Data, such as audio-recordings, will be stored in a secure location and will be deleted when no longer required by the university.

Interested in participating?
Please email me at paul.galbraith@city.ac.uk – I will then be in touch to schedule a time at your earliest convenience to conduct the study.

If you have any questions, please do not hesitate to contact me at the aforementioned email address.

Thank you for your time,

Paul

Appendix C: Recruitment screener

To confirm participants are suitable for the study, the following questions will be asked prior to booking a session with them:

- Do you have a computer science background? [Occupation/education – probe for details, don't want participants that are likely to understand how recommender systems work from a computer science perspective]
- How many hours of TV do you watch per week? [At least four hours]
- How much of that is on the BBC? [At least two hours of BBC programmes]
- Do you use a computer, tablet, or smartphone? [And how often is it used – at least two hours a week.]
- How often do you use the internet? [At least two hours a week]
- What sites do you visit? [Should mention a range of sites e.g. YouTube, Facebook, Amazon, eBay]

[The idea behind the last three question is to understand how experienced participants are with computers/software – so that any actions they make with the prototype system, is more likely to be a deliberate action and not due to them lacking experience with computers/software]

Appendix D: Post session preparation

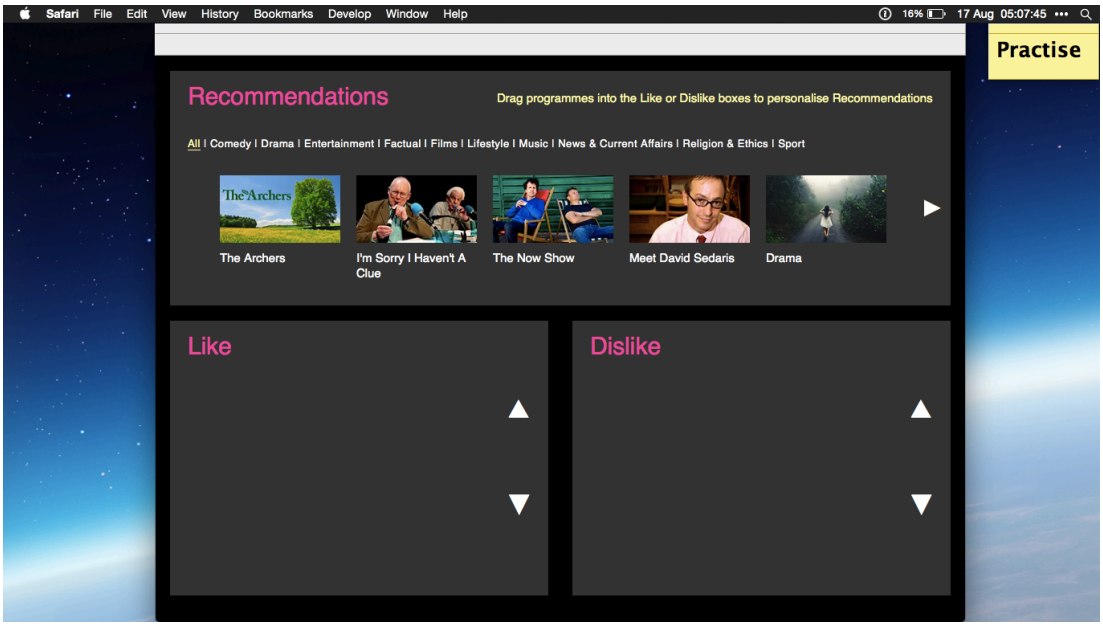
After participant leaves

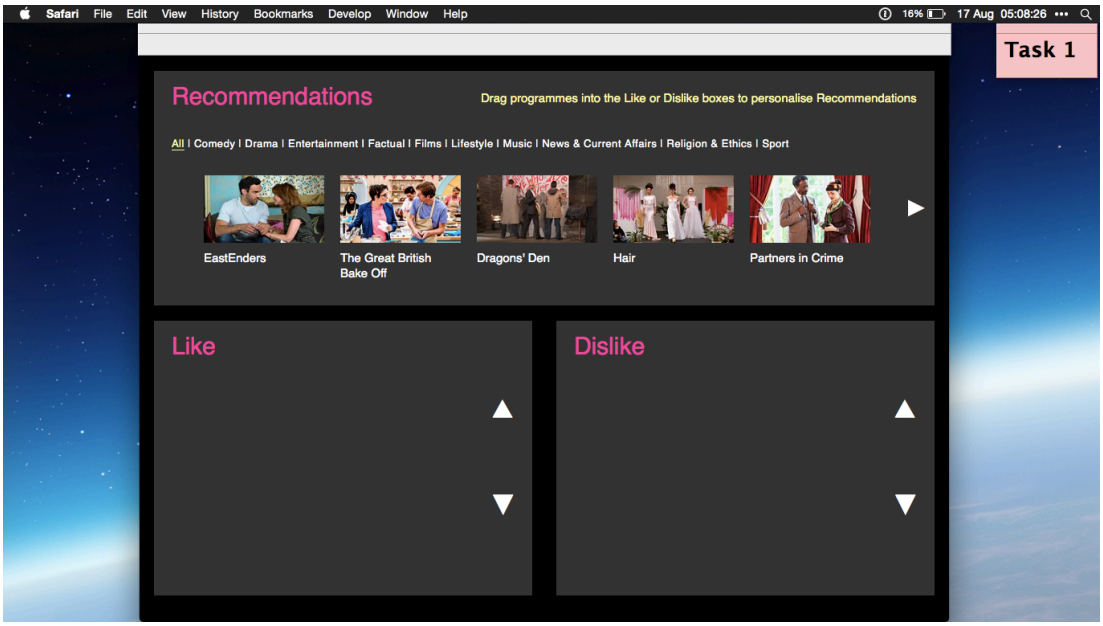
- Back-up both recordings to a flash drive
- Check hard disk space – if less than 5GB remaining, delete files that have been transferred
- Write participant number onto Questionnaires and staple together
- Re-set the Practise and Test Screens
- Prepare forms, questionnaires and Post-its for the next participant

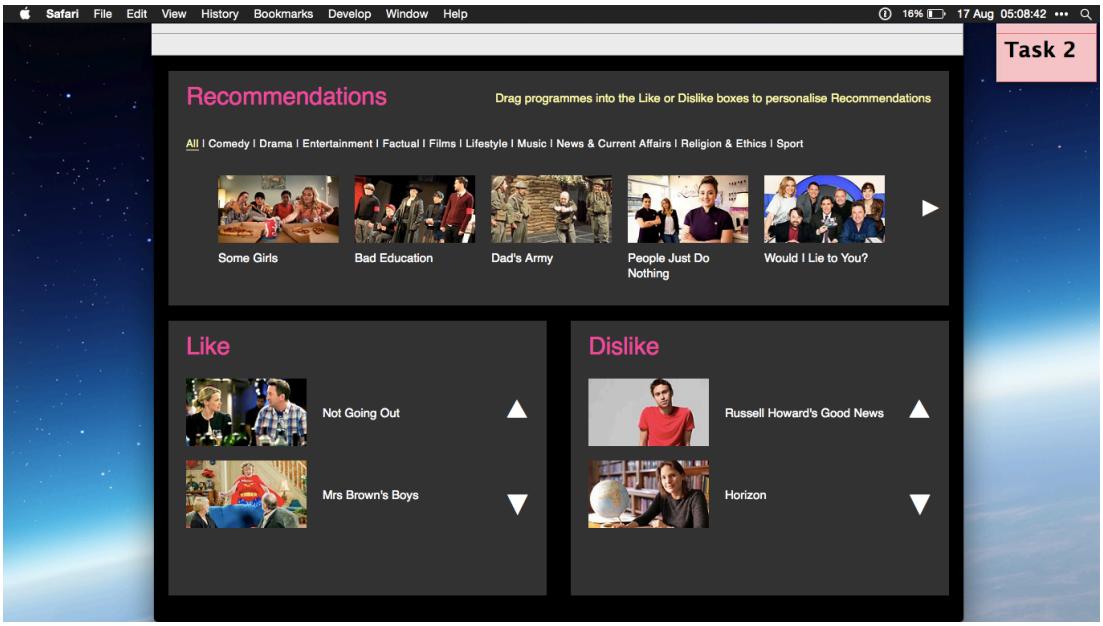
Each evening after sessions

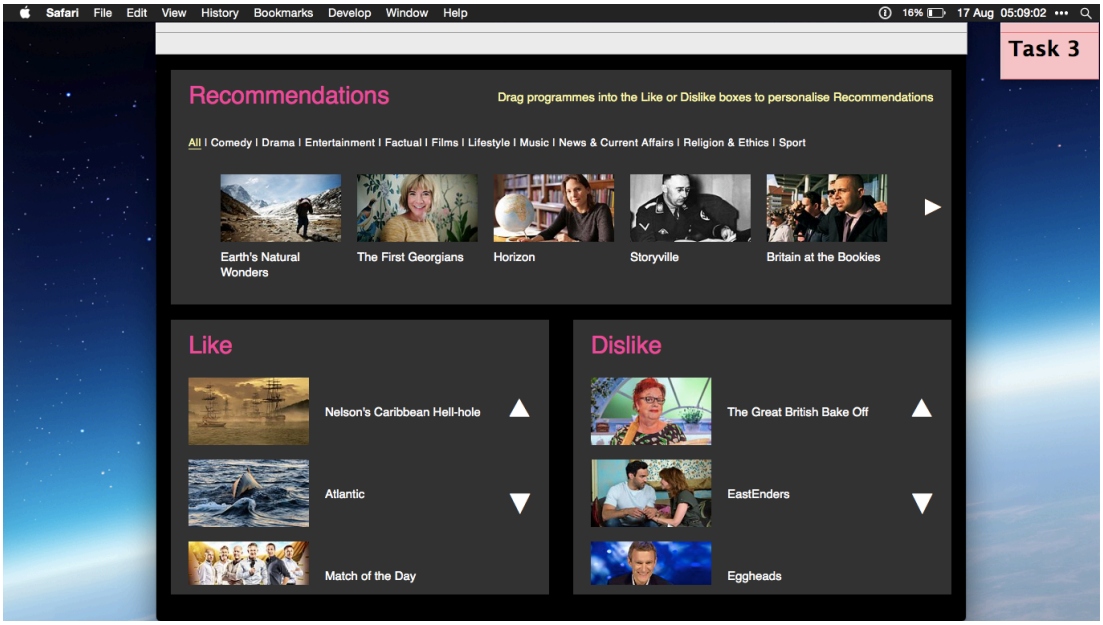
- Transfer both recordings from laptop onto iMac
- Scan Consent Forms and Questionnaires and place originals into secure location
- If time, transcribe sessions

Appendix E: Task setup screenshots






Safari File Edit View History Bookmarks Develop Window Help
16% 17 Aug 05:08:42
Task 2
Recommendations
Drag programmes into the Like or Dislike boxes to personalise Recommendations
All | Comedy | Drama | Entertainment | Factual | Films | Lifestyle | Music | News & Current Affairs | Religion & Ethics | Sport
Some Girls
Bad Education
Dad's Army
People Just Do Nothing
Would I Lie to You?
Like
Not Going Out
Mrs Brown's Boys
Dislike
Russell Howard's Good News
Horizon


Safari File Edit View History Bookmarks Develop Window Help
16% 17 Aug 05:09:02
Task 3
Recommendations
Drag programmes into the Like or Dislike boxes to personalise Recommendations
All | Comedy | Drama | Entertainment | Factual | Films | Lifestyle | Music | News & Current Affairs | Religion & Ethics | Sport
Earth's Natural Wonders
The First Georgians
Horizon
Storyville
Britain at the Bookies
Like
Nelson's Caribbean Hell-hole
Atlantic
Match of the Day
Dislike
The Great British Bake Off
EastEnders
Eggheads

Appendix F: Test script

[Give participant Participant Information Sheet to read]
[Give participant Consent Form to read, initial and sign]

Thank you for agreeing to participate. The session will last up to one hour. If you need to take a break at any point just let me know.

[Give participant Self-Efficacy Questionnaire to complete]

[Show participant Practice Screen]
[Point to all elements on screen as they are being discussed]

Here is the recommender system that you will be using for the tasks I will be asking you to complete shortly. However, this version contains radio programmes, whereas the version you'll be using for the tasks will feature TV programmes.

The system's interface is made up of three boxes:

- At the top is the Recommendations box, containing a row of recommended TV programmes
- Bottom left is the Like box
- Bottom right is the Dislike box

The idea is to drag programmes into the Like or Dislike boxes to personalise Recommendations. So when you drag a programme from the Recommendations box into the Like or Dislike box, it causes the row of recommendations at the top to update according to your selection.

The Recommendations at the top are currently set to display All, alongside are various genres that can be selected instead.

I will now give you a minute to try the system – this is just for practice, so feel free to drag programmes into the Like and Dislike boxes, and to click on the different genres.

Please use the mouse throughout the tasks, not the trackpad.

[Let participant spend a minute using system]

[Start QuickTime Screen Recording (+ audio) – leave running for all tasks]
[Stay on Practice Screen – read this page to participant]

Introduction
With all the TV channels now available, including terrestrial, satellite and online services – there's a vast amount to choose from. This can make selecting a programme to watch from all that's out there quite difficult. Especially as you don't want to waste your time choosing a programme, only to discover it's not for you while watching it.

So when you heard about a new TV programme recommender system, you thought you would give it a try.

Scenario A
Before the system can recommend programmes it believes you may enjoy, you need to teach it your preferences by performing a set-up procedure. This would only be performed once when you first use the system, from then on you would be provided with recommendations based on those initial selections.

Task 1
I would like you to provide the system with enough information about your viewing preferences, so that:

a) **While the 'All' genre is selected;**
b) **The row of recommended programmes at the top, starts with 5 programmes you would like to watch over the next week, that you haven't already seen.**

How you use the system to achieve this is totally up to you, and there's no time limit for the task, so take however long you need. Try to imagine I'm not here while you perform the task, and let me know as soon as you feel you have completed the task.

[Give participant Task 1 Instruction Sheet for reference]

Do you have any questions? Let me know when you're ready to start.

[Switch to Task 1 Screen when participant is ready to start + click safari window]

[After task complete - give participant NASA-TLX 1 to complete]
[Stay on Task 1 Screen - read this page to participant]

Scenario B
It turns out that the recommender system wasn't restricted to a one-time set-up procedure. This means you're free to make changes to the system whenever you wish to, which you've already done.

Your friend David is coming around tonight and he is very particular about what he chooses to watch; often spending half the evening making a decision. So you would like to use the recommender system, to have suggestions ready for what to watch before David arrives. **You know your friend loves history programmes, particularly world history.**

Task 2
I would like you to make changes to the system, though this time **there can only be a maximum of 3 programmes in the Like box and 3 programmes in the Dislike box at any one time, throughout the whole task**, so that:

a) **While the 'All' genre is selected;**
b) **The row of recommended programmes at the top, starts with 5 programmes you believe your friend David would like to watch.**

How you use the system to achieve this is totally up to you, and there's no time limit for the task, so take however long you need. Try to imagine I'm not here while you perform the task, and let me know as soon as you feel you have completed the task.

[Give participant Task 2 Instruction Sheet for reference]
[Add Post-its to screen]

Do you have any questions? Let me know when you're ready to start.

[Switch to Task 2 Screen when participant is ready to start + click safari window]

[After task complete - give participant NASA-TLX 2 to complete]

[Stay on Task 2 Screen - read this page to participant]

Scenario C
Your friend David was very happy with a couple of the recommended programmes the other evening. Due to this success you would like to use the system to prepare a list of programmes for your friend Sandra, who is staying with you this weekend.

Luckily Sandra isn't as fussy as David about what she'll watch, though **she hates anything related to sport, history or nature**.

Task 3
I would like you to make changes to the system, though this time **no programmes should be removed from the Like or Dislike boxes - though there is no limit to how many you put into each**, so that:

a) **While the 'All' genre is selected;**
b) **The row of recommended programmes at the top, starts with 5 programmes you believe your friend Sandra would like to watch.**

How you use the system to achieve this is totally up to you, and there's no time limit for the task, so take however long you need. Try to imagine I'm not here while you perform the task, and let me know as soon as you feel you have completed the task.

[Give participant Task 3 Instruction Sheet for reference]
[Add Post-its to screen]

Do you have any questions? Let me know when you're ready to start.

[Switch to Task 3 Screen when participant is ready to start + click safari window]

[After task complete - give participant NASA-TLX 3 to complete]

[Stop QuickTime Screen Recording – save as PnA]

Appendix G: Interview questions

What I would like you to do now please, is to watch back the three tasks you just completed and talk me through your reasoning for the actions you made. I want to understand how you were expecting the system to react to the choices you made.

You can pause the video at any point, and as often as you need to, so that you can fully explain your reasoning for every action you made.

[Start QuickTime Screen Recording (+ audio) for Retro. Think-aloud]
[Play video of tasks]

[After Retrospective Think-aloud – ask questions below]

1. Please explain to me how you think the recommender system actually works?
2. When you drag a programme into the Like box, what do you think the system does 'under the hood' that causes what it recommends to change?
3. What do you think would happen if the programme Mastermind was dragged into the Like box? [confirm they know what the show is]
4. In hindsight, was there anything you could have done differently to improve the recommendations the system provided?
5. When you first used the system, before any programmes were in the Like or Dislike box, how do you think the recommendations at the top were ordered?
6. Finally, is there anything further you would like to add about your use with the recommender system?

[Stop QuickTime Screen Recording – save as PnB]

Thank you for taking part in my study.

Appendix H: Recommender system reset

Recommender System set-up details

Task 1:
System starts with the 'All' genre selected
Likes: none
Dislikes: none

Task 2:
System starts with the 'All' genre selected
Likes: Not Going Out, Mrs Brown's Boys
Dislikes: Horizon, Russell Howard's Good News

Task 3:
System starts with the 'All' genre selected
Likes: Match of the Day, Atlantic, Nelson's Caribbean Hell-hole
Dislikes: Eggheads, EastEnders, The Great British Bake Off

Appendix I: Information sheet

CITY UNIVERSITY LONDON

PARTICIPANT INFORMATION SHEET

Title of Study: Using interactive recommender systems

I would like to invite you to take part in a research study. Before you decide whether you would like to take part it is important that you understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information.

What is the purpose of the study?
This study is part of a MSc in Human-Centred Systems course and is for my INM363 Individual Project. The project will run for three months from July 2015 to September 2015. It will investigate the use of interactive recommender systems.

Why have I been invited?
You have indicated that you have the domain knowledge for this study, which involves being a regular TV programme viewer, particularly BBC programmes. As well as a regular internet user.

Do I have to take part?
Participation in the project is voluntary, and you can choose not to participate in part or all of the project. You can withdraw at any stage of the project without being penalised or disadvantaged in any way.

It is up to you to decide whether or not to take part. If you do decide to take part, you will be asked to sign a consent form. If you decide to take part, you are still free to withdraw at any time and without giving a reason.

What will happen if I take part?
- The research study will last for three months, though you will only have meet the researcher once for up to one hour (this session)
- You will need to complete four questionnaires
- There will be an observation consisting of three tasks to complete
- A retrospective-thinkaloud will take place after the observation
- An interview will conclude the session

What do I have to do?
You will meet the researcher on one occasion, for up to one hour. You will need to complete four questionnaires. There will be an observation where you will be given three tasks to complete. A retrospective-thinkaloud will take place after the observation. An interview will conclude the session.

What are the possible disadvantages and risks of taking part?
No particular risks have been identified, but please let me know if you feel uncomfortable at any time or wish to stop the study.

What are the possible benefits of taking part?
Although there are no direct benefits to you taking part, I hope you will find participating interesting.

What will happen when the research study stops?
Findings from the study will be written up as INM363 Individual Project. Data will be stored securely, and your identifiable information will be kept separate from the information you provided.

Appendix J: Consent form

CITY UNIVERSITY LONDON

CONSENT FORM

Title of Study: Using interactive recommender systems

		Please initial boxes
1.	I agree to take part in the above City University London research project. I have had the project explained to me, and I have read the participant information sheet, which I may keep for my records. I understand this will involve: • completing several questionnaires throughout the session • being observed while completing several tasks using an interactive recommender system • being interviewed by the researcher • allowing the whole session to be video and audio-recorded	
2.	This information will be held and processed for the following purpose(s): To allow the researcher named below to complete his INM363 Individual Project I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party. No identifiable personal data will be published. The identifiable data will not be shared with any other organisation.	
3.	I understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdraw at any stage of the project without being penalized or disadvantaged in any way.	
4.	I agree to City University London recording and processing this information about me. I understand that this information will be used only for the purpose(s) set out in this statement and my consent is conditional on the University complying with its duties and obligations under the Data Protection Act 1998.	
5.	I agree to take part in the above study.	

Name of Participant	Signature	Date
Name of Researcher	Signature	Date

When completed, 1 copy for participant; 1 copy for researcher file.

CITY UNIVERSITY LONDON

City University London
Northampton Square
London EC1V 0HB
T: +44 (0)20 7040 www.city.ac.uk

Appendix K: Self-efficacy questionnaire

Background Questions

Gender: Female ☐ Male ☐

Age: _____

Self-Efficacy Questions

This part of the questionnaire asks you about your ability to use an unfamiliar piece of software. Often in our jobs we are told about software packages that are available to make work easier. For the following questions, imagine that you were given a new software package for some aspect of your work. It doesn't matter specifically what this software package does, only that it is intended to make your job easier and that you have never used it before.

The following questions ask you to indicate whether you could use this unfamiliar software package under a variety of conditions. For each of the conditions, please indicate whether you think you would be able to complete the job using the software package. Then, for each condition that you answered "yes", please rate your confidence about your first judgment, by circling a number from 1 to 10, where 1 indicates "Not at all confident", 5 indicates "Moderately confident", and 10 indicates "Totally confident".

For example, consider the following sample item:

I COULD COMPLETE THE JOB USING THE SOFTWARE PACKAGE…

			NOT AT ALL CONFIDENT			MODERATELY CONFIDENT				TOTALLY CONFIDENT
	…if there was someone giving me step by step instructions.	(Yes…) No	1	2	3	4	(5)	6	7	8	9	10

The sample response shows that the individual felt he or she could complete the job using the software with step by step instructions (YES is circled), and was moderately confident that he or she could do so.

I COULD COMPLETE THE JOB USING THE SOFTWARE PACKAGE...

			NOT AT ALL CONFIDENT				MODERATELY CONFIDENT				TOTALLY CONFIDENT	
1	...if there was no one around to tell me what to do as I go.	Yes... No	1	2	3	4	5	6	7	8	9	10
2	...if I had never used a package like it before.	Yes... No	1	2	3	4	5	6	7	8	9	10
3	...if I had only the software manuals for reference.	Yes... No	1	2	3	4	5	6	7	8	9	10
4	...if I had seen someone else using it before trying it myself.	Yes... No	1	2	3	4	5	6	7	8	9	10
5	...if I could call someone for help if I got stuck.	Yes... No	1	2	3	4	5	6	7	8	9	10
6	...if someone else had helped me get started.	Yes... No	1	2	3	4	5	6	7	8	9	10
7	...if I had a lot of time to complete the job for which the software was provided.	Yes... No	1	2	3	4	5	6	7	8	9	10
8	...if I had just the built-in help facility for assistance.	Yes... No	1	2	3	4	5	6	7	8	9	10
9	... if someone showed me how to do it first.	Yes... No	1	2	3	4	5	6	7	8	9	10
10	...if I had used similar packages before this one to do the same job.	Yes... No	1	2	3	4	5	6	7	8	9	10

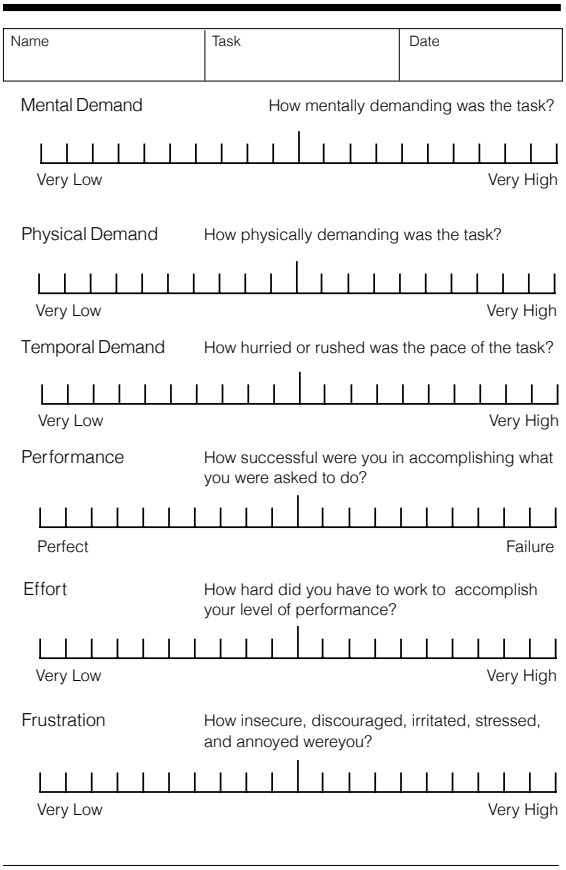
This is the measure of self-efficacy that was used in: Compeau, D. R., & Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. MIS Quarterly. Volume 19, Number 2, pp. 189-211.

Appendix L: NASA-TLX Questionnaire

Figure 8.6

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.



Appendix M: Task instructions

Scenario A
Before the system can recommend programmes it believes you may enjoy, you need to teach it your preferences by performing a set-up procedure. This would only be performed once when you first use the system, from then on you would be provided with recommendations based on those initial selections.

Task 1
I would like you to provide the system with enough information about your viewing preferences, so that:

a) While the 'All' genre is selected;
b) The row of recommended programmes starts with 5 programmes you would like to watch over the next week, that you haven't already seen.

How you use the system to achieve this is totally up to you, and there's no time limit for the task, so take however long you need. Try to imagine I'm not here while you perform the task, and let me know as soon as you feel you have completed the task.

Do you have any questions?

Let me know when you're ready to start.

Scenario B
It turns out that the recommender system wasn't restricted to a one-time set-up procedure. This means you're free to make changes to the system whenever you wish to, which you've already done.

Your friend David is coming around tonight and he is very particular about what he chooses to watch; often spending half the evening making a decision. So you would like to use the recommender system, to have suggestions ready for what to watch before David arrives. You know your friend loves history programmes, particularly world history.

Task 2
I would like you to make changes to the system, though this time there can only be a maximum of 3 programmes in the Like box and 3 programmes in the Dislike box at any one time, so that:

a) While the 'All' genre is selected;
b) The row of recommended programmes starts with 5 programmes you believe your friend David would like to watch.

How you use the system to achieve this is totally up to you, and there's no time limit for the task, so take however long you need. Try to imagine I'm not here while you perform the task, and let me know as soon as you feel you have completed the task.

Do you have any questions?

Let me know when you're ready to start.

Scenario C
Your friend David was happy with a couple of the recommended programmes the other evening. Due to this success you would like to use the system to prepare a list of programmes for your friend Sandra, who is staying with you this weekend.

Luckily Sandra isn't as fussy as David about what she'll watch, though she hates anything related to sport, history or nature.

Task 3
I would like you to make changes to the system, though this time no programmes should be removed from the Like or Dislike boxes, so that:

a) While the 'All' genre is selected;
b) The row of recommended programmes starts with 5 programmes you believe your friend Sandra would like to watch.

How you use the system to achieve this is totally up to you, and there's no time limit for the task, so take however long you need. Try to imagine I'm not here while you perform the task, and let me know as soon as you feel you have completed the task.

Do you have any questions?

Let me know when you're ready to start.

Appendix N: Task frequency sheet

Participant No.		
Task No.		Start time:
Time to first action		
First action (add, remove, switch, scroll, click)		
Time to add/remove first programme to/from Like/Dislike box		
When in All, number of times programmes were scrolled		
Number of different genres clicked (out of 10)		
When in a genre (excluding All), number of times programmes were scrolled		
Number of times genres were clicked (excluding All)		
Number of times All was clicked		
Number of programmes added to Like box		
Number of programmes removed from Like box		
Number of programmes added to Dislike box		
Number of programmes removed from Dislike box		
Number of pauses (of 3 seconds or more)		
Total task time		
Task Completed		
Like box scrolled		
Dislike box scrolled		