spacebook-project.eu

D6.2.2: Final Evaluation of SPACEBOOK Interactive Search Capabilities

Anna Dickinson, Fiona Wragg, Tiphaine Dalmas, Srinivasan Janarthanam, Xingkun Liu, Phil Bartie
Oliver Lemon, William Mackaness, Bonnie Webber

Distribution: Public

SpaceBook
Spatial & Personal Adaptive Communication Environment: Behaviours & Objects & Operations
& Knowledge

270019 Deliverable 6.2.2
30th April 2014

Project funded by the European Community
under the Seventh Framework Programme for Research and Technological Development

The deliverable identification sheet is to be found on the reverse of this page

Project ref. no.	270019
Project acronym	SpaceBook
Project full title	Spatial and Personal Adaptive Communication Environment: Behaviours & Objects & Operations & Knowledge
Instrument	STREP
Thematic Priority	Cognitive Systems, Interaction and Robotics
Start date/ duration	01 March 2011 / 36 Months

Security	Public
Contractual date of delivery M36	M36 = 28 February 2014
Actual date of delivery	30 April 2014
Deliverable number	6.2.2
Deliverable title	Final Evaluation of SPACEBOOK
Type	Report
Status and version	Final 1.0
Number of pages	6
Contributing WP	UEDIN
WP/ Task responsible	HWU
Other contributors	Anna Dickinson, Fiona Wragg, Tiphaine Dalmas, Srini Janarthanam, Oliver Lemon, William Mackaness
Authors	
EC Project Officer	Franco Mastroddi
Keywords	System evaluation; user evaluation; navigation; information retrieval

TABLE OF CONTENTS

Summary 5

1.0 Introduction 6

1.1 Context / Justification 6

Purpose: Evaluating the differences between three different systems I, II and III. 6

2.0 Methodology 9

2.1 Route Selection 9

2.2 Materials 10

2.3 Experimental Protocol 11

2.4 User tasks 14

2.5 Data collected 16

2.6 Data Analysis 18

3.0 Results 19

3.1 Questionnaire Analysis 19

3.2 Direct System comparisons 41

3.3 Task and Landscape 43

3.4 Movement Data 46

3.5 Behavioural Coding 49

3.6 Input processing: speech recognition and understanding 51

3.7 Usability 59

4.0 Key Findings and recommendations 78

4.1 Key Findings 78

4.2 Recommendations 79

5.0 Conclusions 80

References 85

Appendices 86

Appendix 1: Methodology - Task routes and Questionnaires 86

A1.1 *En route* navigation task questionnaire 86

A1.2 *En route* discovery task questionnaire 86

A1.3 Post-experimental debriefing questionnaire 87

A1.4 Navigation Task Maps 88

Appendix 2: Images of landmarks, vistas and objects of interest within the user tasks 91

Appendix 3: Results Tables 94

A3.1 Item Response Frequency Distributions 94

A3.2 Item Response - System association analysis 95

A3.3 System and Age effects on Item Reponse (GLM ordinal regression analysis results) 96

A3.4 System and Age effects on Scale Response (Repeated Measures ANOVA results) 97

A3.5 Task effects on Scale Response 98

SUMMARY

The evaluation of the SpaceBook system presented here, which was carried out in 2013, was designed to assess the functionality of two particular system elements: the Visibility Engine and the multi-threading mode of the Interaction Manager. To this end, three different versions of the system were created, with these two elements either both present (System I) or singly functional (Systems II - visibility engine on, no multi-threading; and III - multi-threaded interaction manager and no visibility engine). Forty-two volunteers (24 younger adults and 18 people over 50) were recruited to test the system versions over a series of navigation and exploratory tasks.Quantitative data on system performance and user behaviour, and qualitative and observational data on user responses and behaviours were collected and analysed.

The quantitative results show that there is a general trend for users to prefer systems II and III over system I, although these are not statistically significant. This evaluation also highlighted the importance of user expectations and variety of user behaviours and responses, dependent on demographics. The critical nature and weaknesses of the system's input processing modules were exposed, and areas for further development were established.

1.0 INTRODUCTION

The concept of the SpaceBook system is that it is, from the user's perspective, a concealed technology; calm and invisible (Weisser and Brown 1995). Therefore, it has been developed as a dialogue-only system, where interaction occurs through spoken commands and queries only, and the system responds with speech. It is intentional that the user is unable to interact in any other way.

The SpaceBook project involved design and implementation of a working prototype. The different components of SpaceBook were brought together, and variously assessed. Following this, there was a requirement to evaluate the whole system, in the field. This report presents that evaluation, carried out in 2013, and its key findings.

1.1 CONTEXT / JUSTIFICATION

Purpose: Evaluating the differences between three different systems I, II and III.

SpaceBook comprises various components: a visibility engine, integrated with a city model, a pedestrian tracker, an interaction manager, and question and answer functionality. Various components are required to convert human utterances into text, to interpret that text, generate a response, and then convert test to speech which is then delivered back to the user (Figure 1.01).

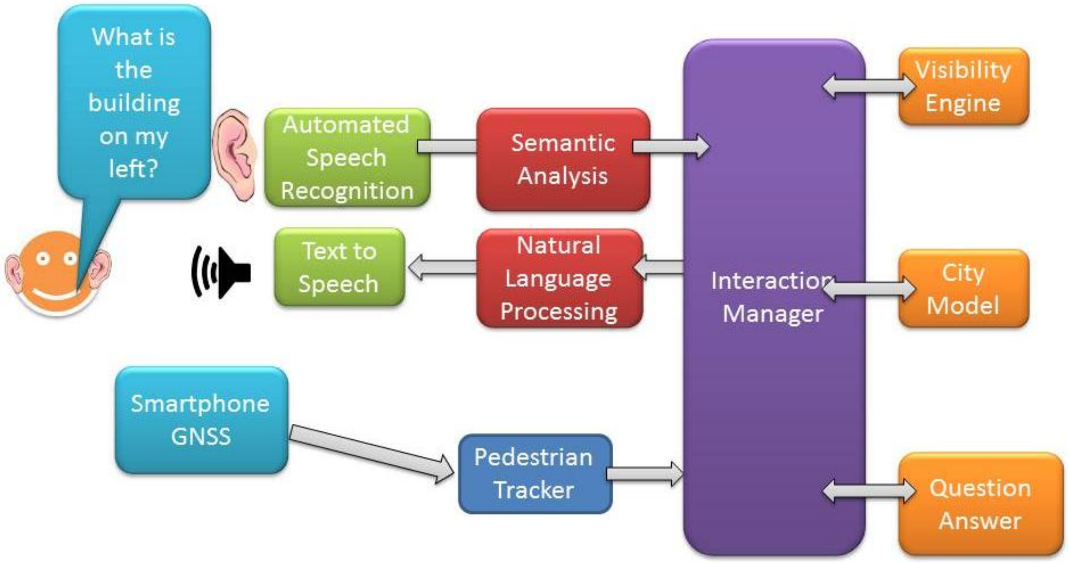


Figure 1.01: The main components comprising the SpaceBook project

Each thread manager will be fed with the input from the user and the dialogue actions generated will be stored in separate queues. This is what we call *multi-queuing*. This approach allows the interaction manager to produce several dialogue actions at the same time although for different conversational tasks.

The queues can be assigned priorities that decide the order in which items from the queues will be delivered to the user. The dialogue actions in the queues are pushed to the user based on an order of priority (see below). This priority can either be fixed or dynamic based on context.

The dialogue actions are generated and stored in queues. Therefore, there is a difference between the time they are generated and time that they are pushed. Therefore dialogue actions in the queues are revised periodically to reflect changes in context. Obsolete dialogue actions will have to be removed for two reasons. Firstly, pushing them to the user may make the conversation incoherent because the system may be speaking about an entity that is no longer relevant and secondly, these obsolete dialogue actions may delay other important dialogue actions from being pushed in a timely manner. In addition, it may also be useful to edit the dialogue actions to include discourse markers to signify topic change (Yang et al., 2008) and bridge phrases to reintroduce a previous topic.

The SpaceBook IM manages the conversation using five conversational threads using dedicated task managers. Three threads: 'navigation', 'question answering' and 'PoI pushing', represent the core tasks of our system. In addition, for handling the issues in dialogue management, we introduce two threads: 'dialogue control' and 'request response'. These different threads represent the state of different dimensions of the user-system conversation that need to interleave with each other components coherently.

Priority is assigned to the above dialogue threads as follows:

Priority 1. Dialogue control (repeat request, clarifications, etc)

Priority 2. Responding to user requests

Priority 3. System initiated navigation task actions

Priority 4. Responses to User-initiated QA actions

Priority 5. PoI Push actions

For instance, informing the user of a point of interest (PoI) could be delayed if the user needs to be given an instruction to turn at the junction they are approaching. For more details on the multi-threaded interaction manager, please refer to Janarthanam and Lemon (2014).

VISIBILITY ENGINE

The Visibility Engine (VE) identifies the entities that are in the user's vista space (Montello 1997). To do this it accesses a digital surface model, sourced from LiDAR, which is a 2.5D representation of the city including buildings, vegetation, and land surface elevation. The VE uses this dataset to offer a number of services, such as determining the line of sight from the observer to nominated points (e.g. which junctions are visible), and determining which entities within the City Model (CM) are visible and highly salient for use as landmarks in navigation instructions, determining visible entities for PoI push. A range of visual metrics are available to describe the visibility of entities, such as the field of view occupied, vertical extent visible, the facade area in view. These metrics are used by the interaction manager and the utterance generator to generate effective navigation instructions. For use as landmarks in navigation instructions, entities such as chain stores (e.g. KFC, McDonalds, Tesco, etc) that are easily identifiable by users are considered as well as buildings and statues. On the other hand, for the PoI push task, we can determine visually interesting buildings by using crowd sourced data. FlickR2 gives us an indication of those features which are considered visually interesting, since people rarely photograph mundane items, but instead photograph unusual or eye-catching buildings. They do however photograph the more interesting and extra-ordinary objects at either end of the beauty spectrum. In addition we use FourSquare data to determine which locations are popular in the city and frequently visited (e.g. cafes, bookshops, bed and breakfast hotels, pubs, *etc*.). For more information on the Visibility Engine, please refer to Bartie *et al.* (2013).

SYSTEM VARIANTS

SYSTEM I:

The first variant, System I, was the full system that features a Multi-threaded Interaction Manager and uses the information from the Visibility Engine. The information from the visibility engine was used for both navigation and information pushing.

SYSTEM II:

System II used a single-threaded interaction manager, comparable to the lean, thin SpaceBook system from Year 1. The IM had one queue into which the dialogue actions from various tasks were pushed. These were popped out one after another to be presented to the user. In contrast to System I, it cannot prioritise the dialogue actions in the queue that were to be delivered.

SYSTEM III:

System III used the multi-threaded interaction manager but the visibility engine was turned off. Therefore the navigation and information pushes used only proximity information to identify landmarks and points-of-interest.

We expect that System I will perform better compared to the two other systems due to the complex nature of the interaction. With the visibility engine identifying a large number of points-of-interest to present for the information pushing task, the multi-threaded IM will have an advantage as it segregates them in a separate queue and pushes them with lower priority compared to other, more critical tasks such as a request from the user (user pull). This means that when the user asks a question, it will be answered before the PoI information in the queue is pushed. This compares to the single-threaded IM, using the single queue, which has to first push all the content out before it can answer the user's query. We expect users to be more satisfied with System I as it allows more control (users' requests are given a high priority) and more coherence (dialogue acts are not incoherently mixed up in a single queue).

2.0 METHODOLOGY

SpaceBook is unique and there are no directly comparable technologies. Firstly it is unique in providing information based on what is in the field of view (not just what is proximal); second it handles a mix of information, some specific to wayfinding, and some relating to the description of its surroundings; third, SpaceBook responds to the user's requests for further information on their surroundings.

The closest comparator system, Siri, is not 'spatially aware' and cannot describe things in the field of view (this is because it does not have a viewshed model). Thus in the absence of a geographical context, it cannot answer deictic questions such as 'what is that?'

Most other wayfinding technologies are either 1) ocular centric requiring interaction with the device via keystrokes and screen pinching, or 2) use haptics such as vibration or abstract sound (Brewster, 2002) SpaceBook is solely dialogue based, concerned with mediating different types of information to support exploration. The challenge lies not in developing wayfinding software, nor QandA methodologies, but in integrating these technologies and utilising geographic information as a contextualising focus so the user is free to explore the city.

In light of this it was decided to isolate various components of the system and assess the impact of that diminished functionality upon the user experience across a range of tasks and routes The experiment was intended to explore the usefulness of (a) the Visibility Engine, and (b) multi-threaded dialogue, and to contrast the updated system with a basic system, equivalent to the version tested in 2013.

2.1 ROUTE SELECTION

The experiment comprised three legs. The legs were chosen for their diversity of views and route complexity. The routes were co-located to facilitate the running of the experiments (Figure 2.01, 2.02).

At the Black Watch Memorial and St Giles Cathedral, users also completed Discovery tasks, using the exploration functionality of SpaceBook to find out about the area they were in.

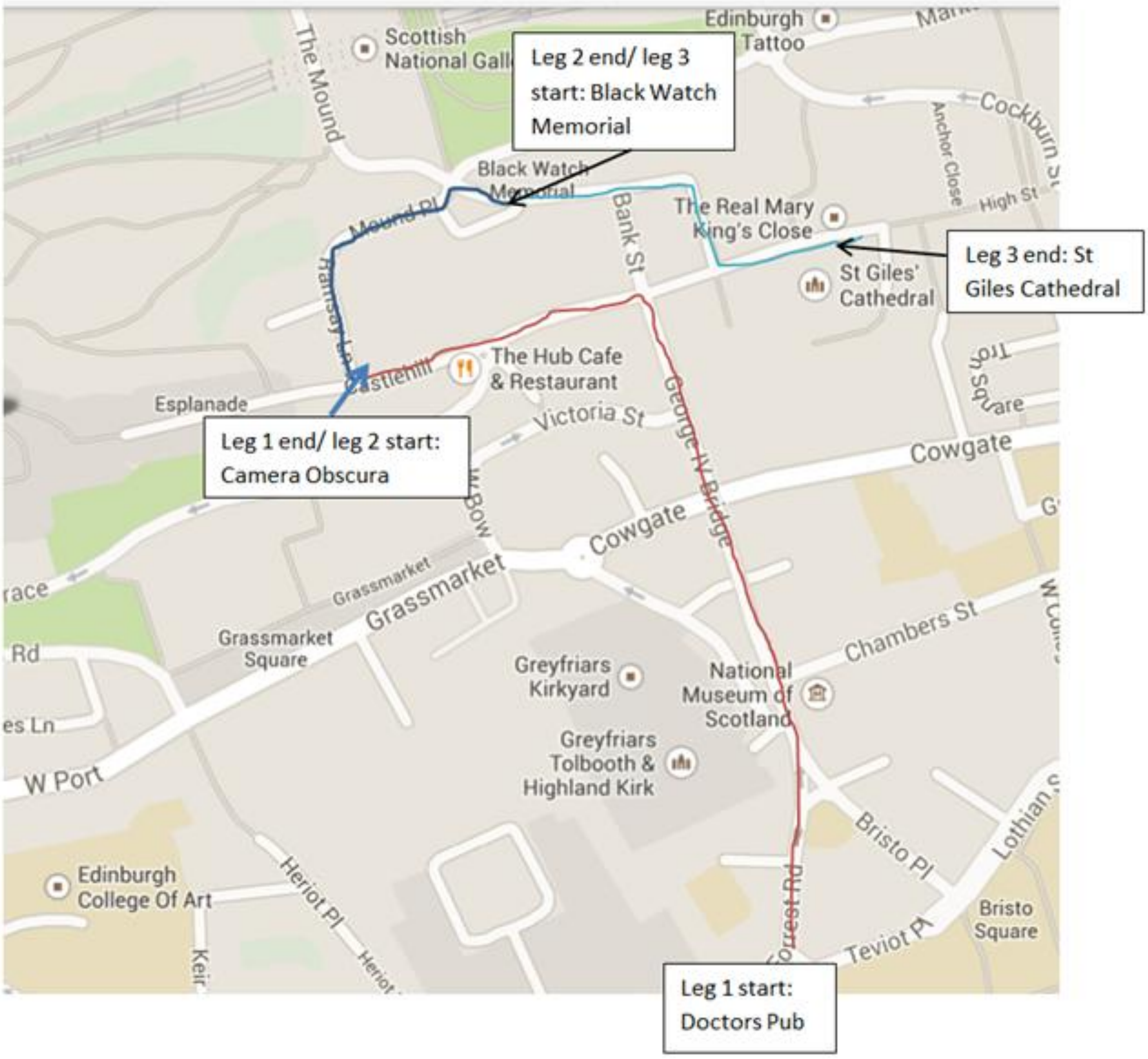


Figure 2.01: The four routes in the city of Edinburgh.

See Appendix 2 for a range of key landmarks that users were guided by, and vistas and objects of interest that users passed.

2.2 MATERIALS

HARDWARE

The user interacted with the SpaceBook system via a telephone call made from a Samsung Galaxy SIII on which the SpaceBook connection was installed as an application. The handset was worn by the user in an armholder, and an earphone/microphone headset was connected to the handset.

QUESTIONNAIRE

At the end of each segment of the study (see below for study structure), participants completed a ten-question questionnaire (see Appendix 1 for full questionnaire). Questions were grouped in three sections:

Interaction Ease (Questions: "SpaceBook understood me", "I found SpaceBook easy to understand", "I felt as if SpaceBook was responding to my questions", and "I was always sure what SpaceBook was talking about") was a measure of subjects' comfort in interacting with the system, and its accessibility from a user's perspective. It was anticipated that these questions would discriminate between System I and System II, because of the way in which the multi-threaded dialogue affects the responses to users' questions.

Information Content (Questions: "Landmarks that SpaceBook used to direct me were obvious and useful" and " I found what SpaceBook said interesting". Additionally, in some tasks, the question "SpaceBook gave me the right amount of information" was used). These questions were asked in order to distinguish between the systems in both circumstances where the user is pulling information from the system, and when the system is pushing information at the user.

Control and Confidence (Questions: "I found the task easy", "I felt in control of my journey" and "I felt confident I was going to reach my destination"). We anticipated that these questions would show differences between the systems with the Visibility Engine (Systems I and II), and the system without (System III).

2.3 EXPERIMENTAL PROTOCOL

TIMING AND LOCATION OF EXPERIMENT

The Evaluation was carried out in Edinburgh in October - December 2013, with volunteer participants recruited from the local population. At this time of year, Edinburgh's streets are relatively quiet, compared to Christmas and summer seasons, when visitor numbers can be very high. This meant that testing was not unduly impeded by large crowds of people, which could have an effect on user movement patterns and ability to interact with the system, or by temporary street closures and re-routing.

PARTICIPANTS

42 participant were recruited: 24 students (8 male/16 female; mean age 23, range:16-40) and 18 people over 50 (10 male/ 8 female; mean age 62.4, range: 52-76).

Students were recruited through a careers service email and an event-booking system (eventbrite); older adults were recruited through local organisations, including the University of the Third Age, and the "Get Up and Go!" Programme for over-50s. All participants rated themselves "Fit and able", and could walk for 90 minutes and cope with steep and uneven ground; they were all native speakers of English, with a range of accents (including Northern Irish, New Zealand, and Indian).

PROCEDURE

Participants attended for a two hour session and were paid £25 on completion of the experiment.

PRE-TEST BRIEFING: The session started at the Informatics Forum where the participant met the researcher who explained the study, administered a demographic questionnaire (age, fitness, familiarity with smart phones) and explained the informed consent form. Then the participant and the researcher walked to the start of the route where the researcher fitted the SpaceBook system, and started it, and repeated the experiment instructions, giving the participant the opportunity to ask questions. The participant was then given the first task.

INSTRUCTIONS TO PARTICIPANTS: Participants were told that SpaceBook is a phone version of a tour guide that can help you find places, and tell you about the city as you walk through it, and the idea is that SpaceBook speaks to you, and you can talk to it. They were asked to imagine that they were tourists, exploring Edinburgh, who were in no particular hurry, and that the researcher would give them tasks that they were to complete as independently as possible. SpaceBook would tell them about things that it thought they might be interested in, and they could ask it questions about things that they wanted to know about. When they spoke, SpaceBook would listen, and beep when they had finished to acknowledge that it had heard and was ready to respond. They could interrupt SpaceBook, or stop it delivering information, using the command “stop”.

TASK PROTOCOL: As participants carried out the tasks, the researcher followed at a distance allowing them to observe, but discouraging unnecessary communication; the participant was instructed to complete the task on their own, using the system, and only talk to the researcher when they were completely stuck. As the system can only record data that is produced directly from the user’s interaction (*e.g.* a question) or the user’s location, certain user behaviours could not be captured by the system. Thus, the researcher observed participant behaviour, where possible, and recorded the following coded behaviours whenever the participant halted in their course:

C	Confusion	Participant looks confused; appears to be listening intently to the system, or looking around them
F	Frustration	Participant appears frustrated
Q	Questions	Participant has paused in order to ask SpaceBook a question
T	Traffic	Passing traffic has forced participant to pause on their route

		before crossing a road
I	Interest	Participant appears to be listening intently to the system and/or looking at their surroundings

This list represents both negative and positive possible user behavioural responses during the experiment.

Tasks were of two types: Navigation tasks, and Discovery tasks. After each navigation task the participant answered a questionnaire (Appendix 1.1), rating their experience of the task and system in terms of ease of use and other measures. After each discovery task, there was a shorter set of questions (Appendix 1.2).

System order was pseudo-randomised to control for order effects, and a latin-square type design of subject treatment and system replication was employed on the navigation tasks (Table 2.01):

		System		
		I	**II**	**III**
Navigation Task	**1**	14	14	14
	2	15	14	14
	3	14	15	14

Table 2.01. Replication design of subjects across systems and tasks

POST-EXPERIMENTAL DEBRIEFING: After all tasks had been completed, the participant completed a final short questionnaire (Appendix 1.3) in which they were asked to directly compare the three systems that had been used on the navigation tasks (via task identification), on a range of ease of use measures.

2.4 USER TASKS

Each participant completed three navigation tasks, each under a different system (Fig. 2.02a-c).

Navigation targets:

1. Doctor's Pub - Camera Obscura
2. Camera Obscura - The Black Watch Monument
3. Black Watch Monument - St Giles' Cathedral

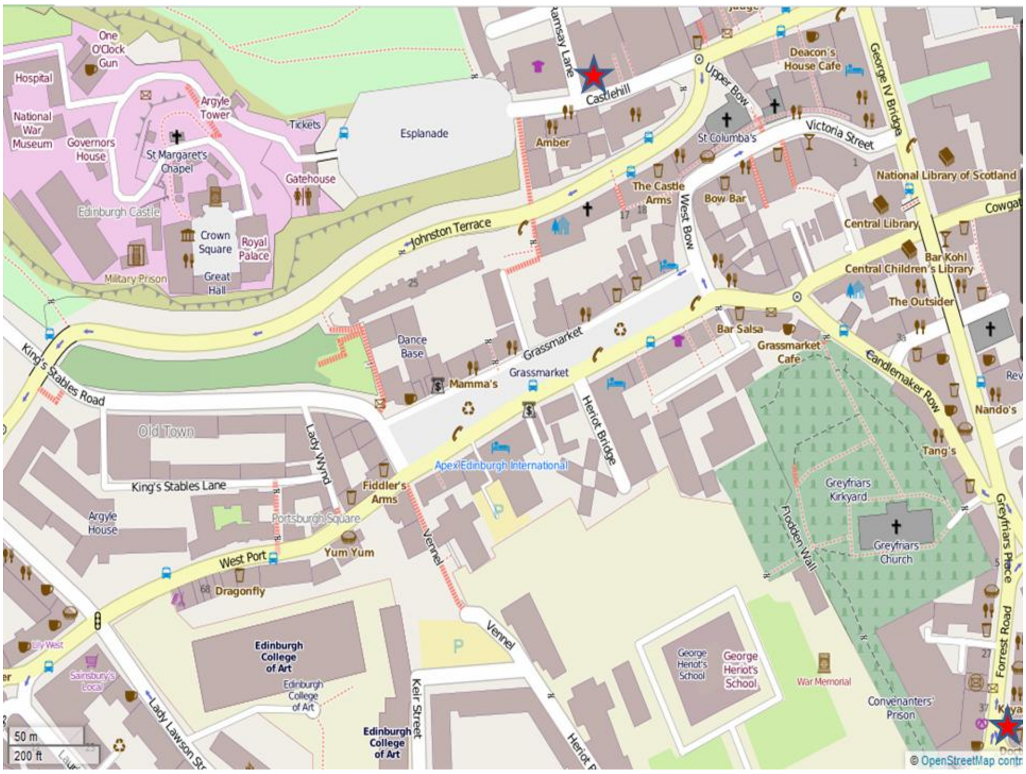


Fig. 2.02a). Navigation Task 1 area map. Red stars indicate start point (Doctor's Pub, Forest Road) and target (Camera Obscura, Castle Hill)

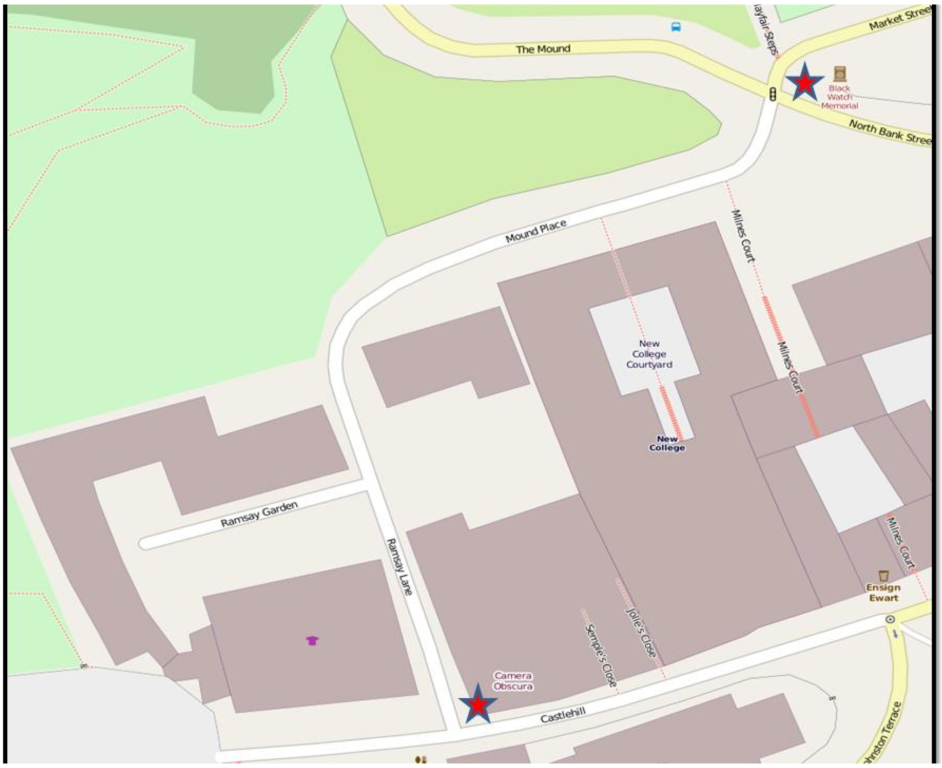


Fig. 2.02b) Navigation Task 2 area map Red stars indicate start point (Camera Obscura, Castle Hill) and target (Black Watch Monument, Market Street)

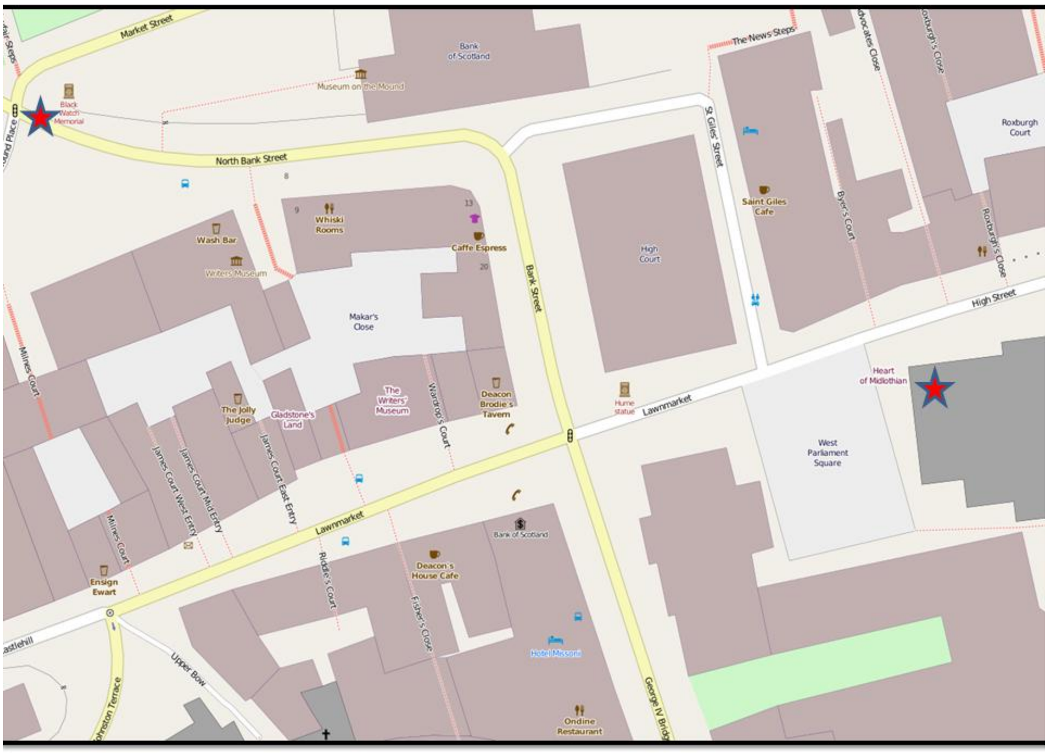


Fig. 2.02c) Navigation Task 3 area map Red stars indicate start point (Black Watch Monument, Market Street) and target (St Giles Cathedral, West Parliament Square)

Each of the following discovery tasks was completed under Systems I or III (ie with or without the Visibility Engine and deictic questions):

VISTA TASK: After the Black Watch Monument task, participants were asked to get the system to tell them about three objects of interest. The task took place on The Mound, overlooking the centre of Edinburgh and including possible targets such as The Scott Monument, the Castle and Jenners.

DISCOVERY TASK: Around St Giles' Cathedral, participants were asked to find another three items of interest. In the area there were many historically significant landmarks, e.g. the Hume statue, Adam Smith's statue, the Mercat Cross, and Mary King's Close.

2.5 DATA COLLECTED

A) SENSOR AND POSITIONAL DATA

The following metrics on participants' movements whilst completing each task were recorded:

- average moving speed
- total task duration
- time spent moving during the task
- percentage of the task spent stationary
- total distance travelled during completion of the task

B) USER / SYSTEM UTTERANCES

As recorded in the System's information database. *i.e.* user speech parsed by the ASR module, and System utterances generated by the NLP module.

C) QUESTIONNAIRE RESPONSES

Questionnaire responses in both qualitative (free text) and quantitative (ordinal categorical response) formats were collected (see appendix A1 for full questionnaire). Categorical questions were presented with a Likert-type response options format, where subjects rated their level of agreement with a given statement on the following ordered scale:

1 - Strongly disagree

2 - Disagree

3 - Somewhat disagree

4 - Neither agree nor disagree

5 - Somewhat agree

6 - Agree

7 - Strongly disagree

For certain questions, a 6-point forced-choice scale was offered, containing no neutral option.

D) AUDIO TRANSCRIPTIONS

All speech audio (plus noise) that entered the headset microphone was recorded in a separate time-stamped audio file, for direct comparison with the time-stamped user utterances recorded by the ASR module.

2.6 DATA ANALYSIS

QUESTIONNAIRE DATA

Categorical responses to each of the questions listed in section 1 were recorded in Likert format, and also converted to a binary format, representing a positive response to the item (agree), and a negative response to the item (disagree). Item (question) responses were analysed separately as ordinal data, and then used to construct scales (through summing) that represented the three key user perspectives that were being assessed: Confidence and Control; Interaction Ease and Information Content; and which were analysed for the effects of system.

POSITIONAL DATA

Data on participants' movements were analysed for effects of System.

ASR DATA

The evaluation transcripts were used to perform a qualitative review of exploration-related utterances. An utterance is considered "exploration-related" when it is either a question about user surroundings or the city in general ("What is this monument?", "Tell me about David Hume"), or an element of clarification pertaining to an exploration dialogue ("No not that one. The small one"). System exploration utterances (pushes) were reviewed, but essentially the focus was on user pulls.

First, transcribed utterances from the audio files were aligned with the ASR output received by the system and the actual incoming/outgoing of each utterance in the system work flow was checked. Transcriptions provided by WP6 also form a valuable corpus of user utterances that draw an outline of common exploration language patterns used by human subjects in a voice-only system. A summary of these patterns with examples from the street evaluation is provided.

3.0 RESULTS

Presented below are, in order, the results of the *en route* user questionnaire analysis (divided into three areas: Confidence and Control, Interaction Ease and Information Content), the post-experimental direct system comparison questions, effects of task and landscape, user movement data, ASR analysis and a qualitative usability report.

3.1 QUESTIONNAIRE ANALYSIS

CONFIDENCE AND CONTROL

The items related to Measures of Confidence and Control were:

- I found the task easy
- I felt in control of my journey
- I felt confident I was going to reach my destination

It was anticipated that systems where the Visibility Engine is on (Systems I and II), should inspire confidence in the user, and increase the sense of control, as the user is being provided with landmarks to navigate by that are further ahead of him or her, and therefore they are not constantly waiting for reassurance that they are on the correct trail.

The results showed as slight preference for system III in terms of perception of task ease, and feeling in control, whereas slightly more people had confidence in finishing the task when working under System II.

Item	**System I**		**System II**		**System III**	
	median	Positive percentage cut-off	median	Positive percentage cut-off	median	Positive percentage cut-off
I found the task easy	5	59.4	5	60.5	5	64.3
I felt in control of my journey	4	48.4	5	51.2	5	54.8
I felt confident I was going to reach my destination	5	68.0	6	69.8	6	66.7

Table 3.01. Descriptive statistics of item responses under different systems, showing median item response scores, and positive percentage cut-off values for each item contributing toward the Confidence and Control Scale. Percentage cut-off values indicate the percentage of subjects who rated a given item positively (*i.e.* 5 "somewhat agree" to 7 "strongly agree"). All items scored on a 7-point Likert-type scale.

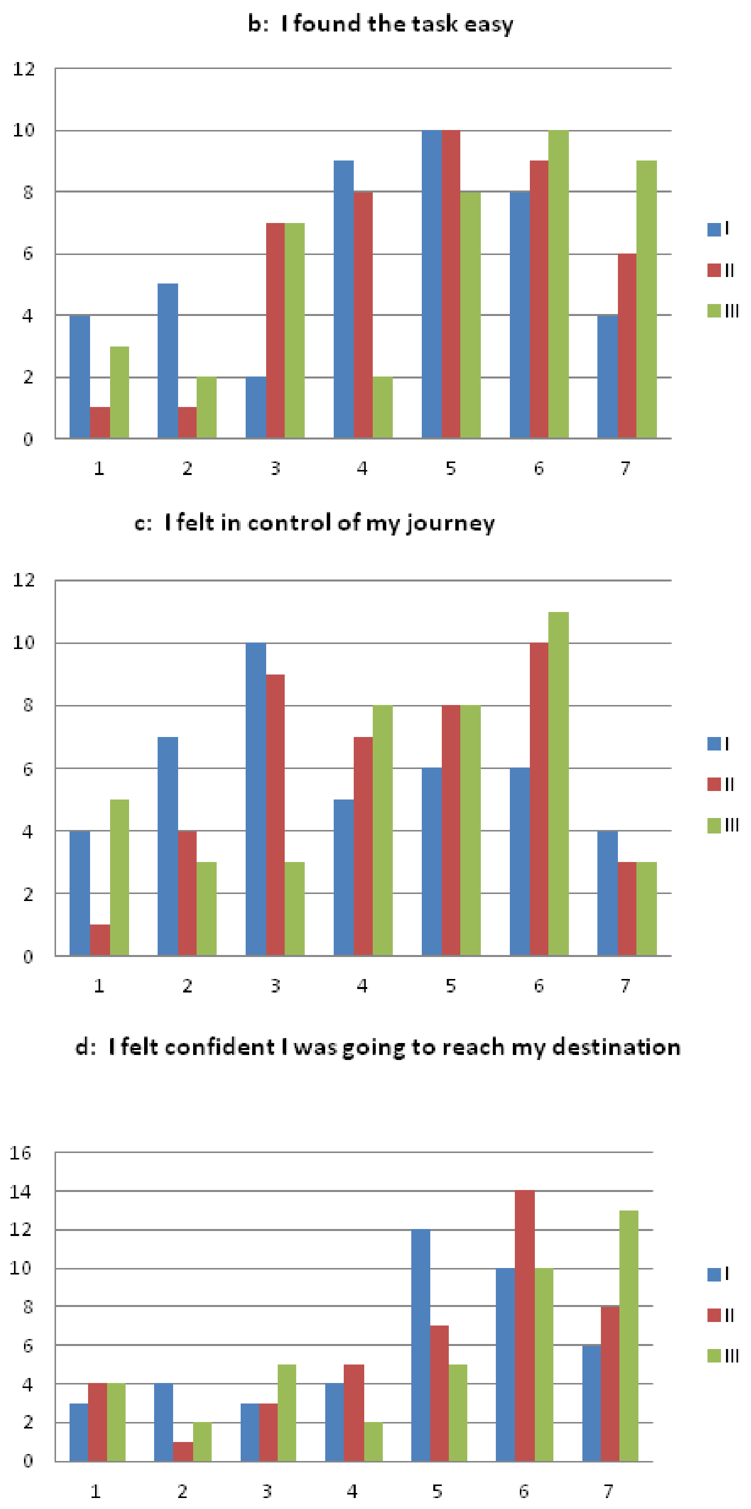


Fig. 3.01. Confidence and Control item response frequency distributions under 3 different versions of the system. Item response categories range from 1 ("strongly disagree") to 7 ("strongly agree").

Table 3.01 suggests that the perception of task ease was largely similar across systems, but that subjects felt less in control, and less confident that they were going to reach the destination when working under System I, compared to Systems II and III.

Item response frequency distributions were significantly different from a normal distribution (Appendix 3.1), necessitating a non-parametric approach for item analysis. After conversion to binary format (agree/disagree), when items were analysed separately for association between response and system, only item c) ("I felt in control of my journey") showed any significant association $\chi^2_{[2]}$=5.054; p=0.042 (see Appendix 3.2 for all results) with significantly more respondents disagreeing with the statement for System I.

Overall, when considering a sense of confidence and control, users frequently reported SpaceBook enhancing this for them, and in particular the regular frequency of location updates and/or information updates:

> "(SpaceBook contributes towards a sense of confidence) especially when it keeps you consistently updated. You always know that it knows where you are." [B02]
>
> "SB kept giving notification of how far I needed to go and that gave me confidence."[B06]
>
> "Distance countdowns build confidence that you are progressing towards your destination."[B08]
>
> "It was almost like a conversation... SB made me feel, the whole time, as though I was being led by a tour guide." [B09a]
>
> "Not just the quality of information, but the frequency affirms where you are "[C14]
>
> "It gave me distance lengths, so I knew how close it should be."[B13]
>
> "SB was very clear where I was going, and the distance countdowns helped."[B16]
>
> "Carry on" would have been good. Often it was silent but I'd have liked some reassurance.[C02]
>
> "Metre countdowns (to target) are good as you then have a general expectation of how far to walk."[C14]
>
> "SB gave me extra information on where I was so that gave me confidence in the system, but that confidence wavers according to SB performance. For task 3 it got my location wrong. Directions which as 'walk ahead' are not useful - you need to know for how far, or until what point is reached. You need back-up or reassurance from SB that I am on the right track."[C08]
>
> "SpaceBook pointed out helpful places along the way, so you knew you were on the right track."[C11]
>
> " It would be good to have some sort of 'cancel' or 'reset' password. Language is important. There are multiple ways of asking a question, so when SB repeats what I've said, it reinforces that feeling of control."[B09a]

The ability to interact with SpaceBook, or lack thereof, also affected users' sense of control:

> "On task 1 (System II) it left silences, for me to talk into; on the other two tasks it didn't leave me much chance to speak so I didn't feel as though I had very much input."[C08]

Additionally, users indicated that trust was an important issue - feeling that there was a connection with the system:

> It was like being with someone, or walking with a companion, and that helped."[C01; task 3; System III]

"It's more a question of trust than control."[B22]

"I got lost but the system redirected me well, it was playful."[B22]

"I trust the system – instructions were correct, navigation by landmarks; very precise."[B21]

""Carry on" would have been good. Often it was silent but I'd have liked some reassurance.."[C02]

"SB in general made me feel less in control. Being reliant on a computer system can make you feel on edge."[B16]

In order to test the effects of both System and Age, a Generalised Linear Modelling approach was taken, using ordinal regression of each item's responses with System and Age as independent variables, and building the model using a logit link function. Visual examination of cumulative percent distribution graphs (Fig. 3.02) show some differences in likelihood of cumulative scores for each of the three Confidence and Control items, with the likelihood of subjects responding negatively under system I being higher than for systems II and III. There is also a difference between the age groups of subjects: older participants had a greater likelihood of responding negatively than did younger participants. However, none of these differences is statistically significant (Appendix 3.3).

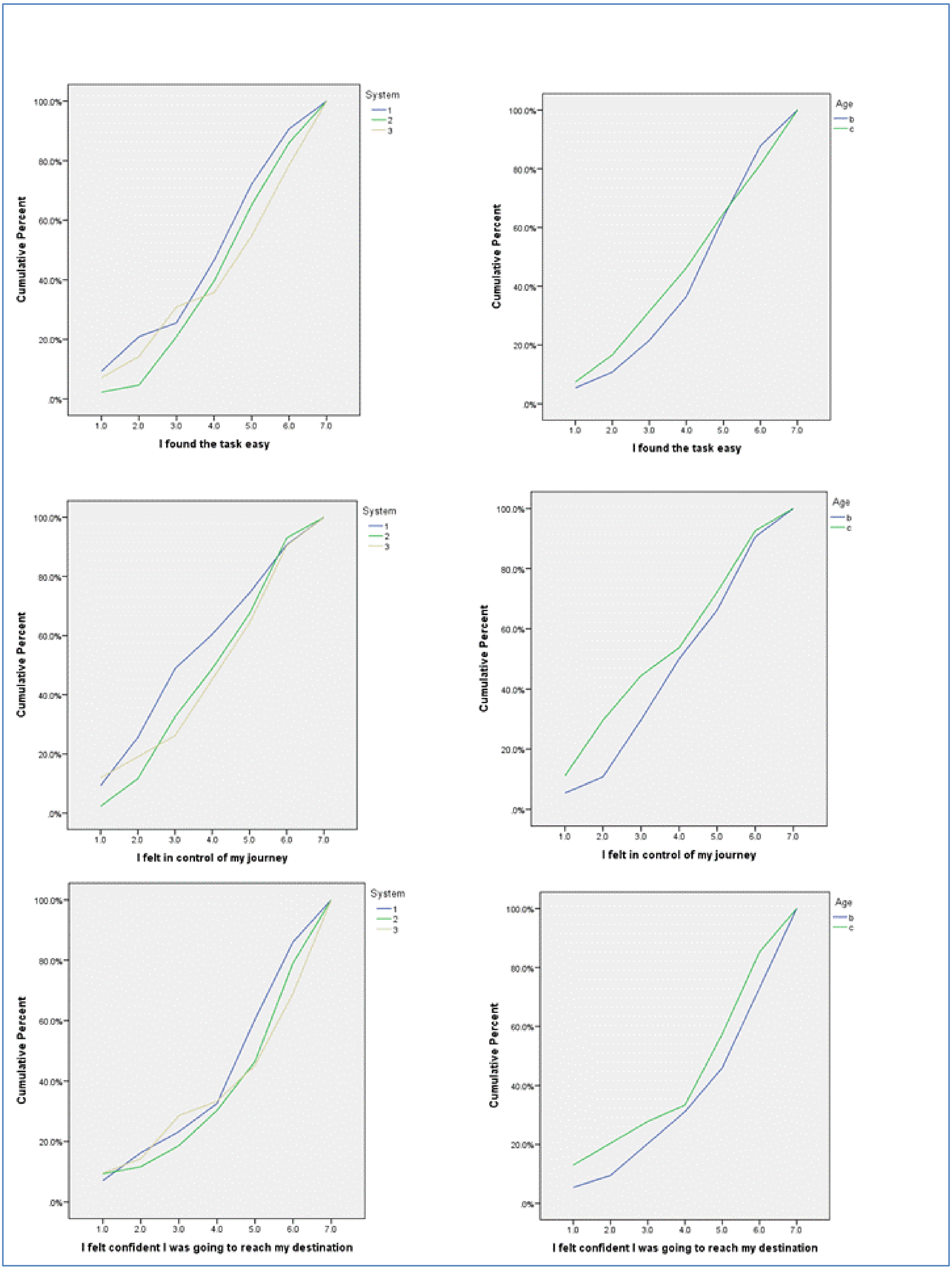


Fig. 3.02. Confidence and Control Scale item response cumulative percent graphs, showing age- and system-effects

Table 3.02 and figures 3.03 and 3.04 show the distributions and descriptive statistics of the Confidence and Control scale under each System.

Scale	System I			System II			System III		
	mean	SD	% 'satisfied'	mean	SD	% 'satisfied'	mean	SD	% 'satisfied'
Confidence and Control	13	4.66	62.8	14.3	4.14	76.7	13.88	5.25	69.8

Table 3.02. Confidence and Control scale response descriptive statistics across systems, showing mean, Standard Deviation, and proportion of respondents scoring satisfied for a given scale, where a satisfied score, x_s, is defined:

$$Xs > \frac{scale\ max}{2}$$

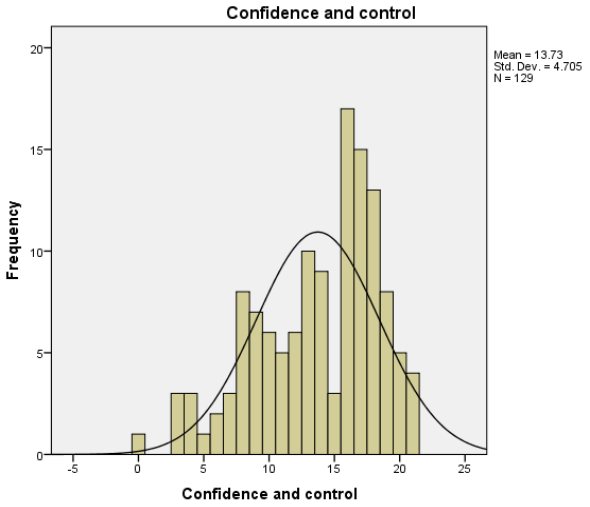


Fig. 3.03. Distribution of total Confidence and Control scale values

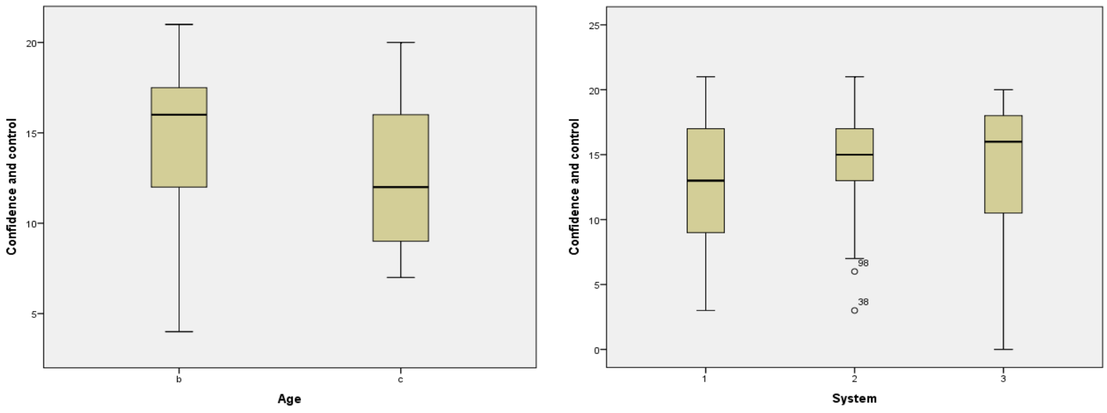


Fig. 3.04. Average confidence and control scale values, for different age groups ('b' = participants aged 18-30; 'c' = participants aged 50yrs+) and System versions.

Descriptive statistics suggest an age effect; older respondents tend to have lower Confidence and Control scores than younger respondents (Fig. 3.04); but there appears to be no real difference between Systems. A General Linear Model, repeated measures design ANOVA, (utilising a one-between, one-within design) was used to investigate the effects of system and age on scale response, after confirming that the model assumptions of sphericity and residual normality were met by the data (see Appendix 3.4 for full results). Neither System ($F_{[2,40]}$=1.304; p=0.283) nor Age ($F_{[1,41]}$=1.174; p=0.285) had any significant effect on Confidence and Control measures, and there was no interaction effect between Age and System ($F_{[2,40]}$=0.311; p=0.734).

Important factors influencing users' sense of control or confidence was the quantity and regularity of information they received from the system, and the accuracy of the system's responses. Any positioning or ASR problems negatively impacted on users' feelings of confidence or control:

> No information was given on whether to cross the street" [C14]
>
> "I was not told to cross the road at almost any point" [C15]
>
> "Non-description of buildings is difficult for tourists" [C15]
>
> "at the junction of Victoria St and George IV bridge it was not giving any directions at all, despite there being four alternative directions."[C10]
>
> "And at the complex junction (by museum of Scotland) much more info is needed to guide user where to go, as that junction is difficult to cross with complex traffic flow."[C04]
>
> "when SB got things wrong it didn't give me much confidence."[C08]
>
> "When SB is talking about things you can't see, or identify, it has an unsettling effect."[C06]

The system's choice and description of landmarks and targets may also have an effect; poorly visible landmarks can confuse or annoy the user:

> "The use of street names instead of landmarks made things difficult."[B12; task3; system I]
>
> "at Milnes Court it was using landmarks that couldn't be seen."[B02; task 2; System III]
>
> "I didn't know what the Black Watch monument looked like, and SpaceBook gave no description, so I didn't know what it was when I got there."[B16]

INTERACTION EASE

The items related to Measures of Interaction Ease were:

- SpaceBook understood me
- I found SpaceBook easy to understand
- I felt as if SpaceBook was responding to my questions
- I was always sure what SpaceBook was talking about

Here, it was anticipated that there should be a difference between systems I and II, and III, based on the Visibility Engine being, respectively, on or off. Additionally, the multi-threading on Systems I and III would be expected to affect accessibility, by providing a system that is more responsive to the needs of the user, and delivers information in a context-appropriate manner.

Item	**System I**		**System II**		**System III**	
	median	Positive percentage cut-off	median	Positive percentage cut-off	median	Positive percentage cut-off
I found SpaceBook easy to understand	6	78.9	6	88.4	6	69.0
SpaceBook understood me	4	45.2	4.5	50.0	4	48.8
I felt as if SpaceBook was responding to my questions**	3.5	37.3	4	37.8	4	39.0
I was always sure what SpaceBook was talking about**	5	53.5	5	57.1	4	47.6

Table 3.03. Descriptive statistics of item responses under different systems, showing median item response scores, and positive percentage cut-off values for each item contributing toward the Interaction Ease Scale. Percentage cut-off values indicate the percentage of subjects who rated a given item positively (*i.e.* 5 "somewhat agree" to 7 "strongly agree" on a 7-point scale, or 4 "somewhat agree" to 6 "strongly agree" on a 6-point scale). The first two items scored on a 7-point Likert-type scale; items marked ** were scored on a 6-point, forced choice scale.

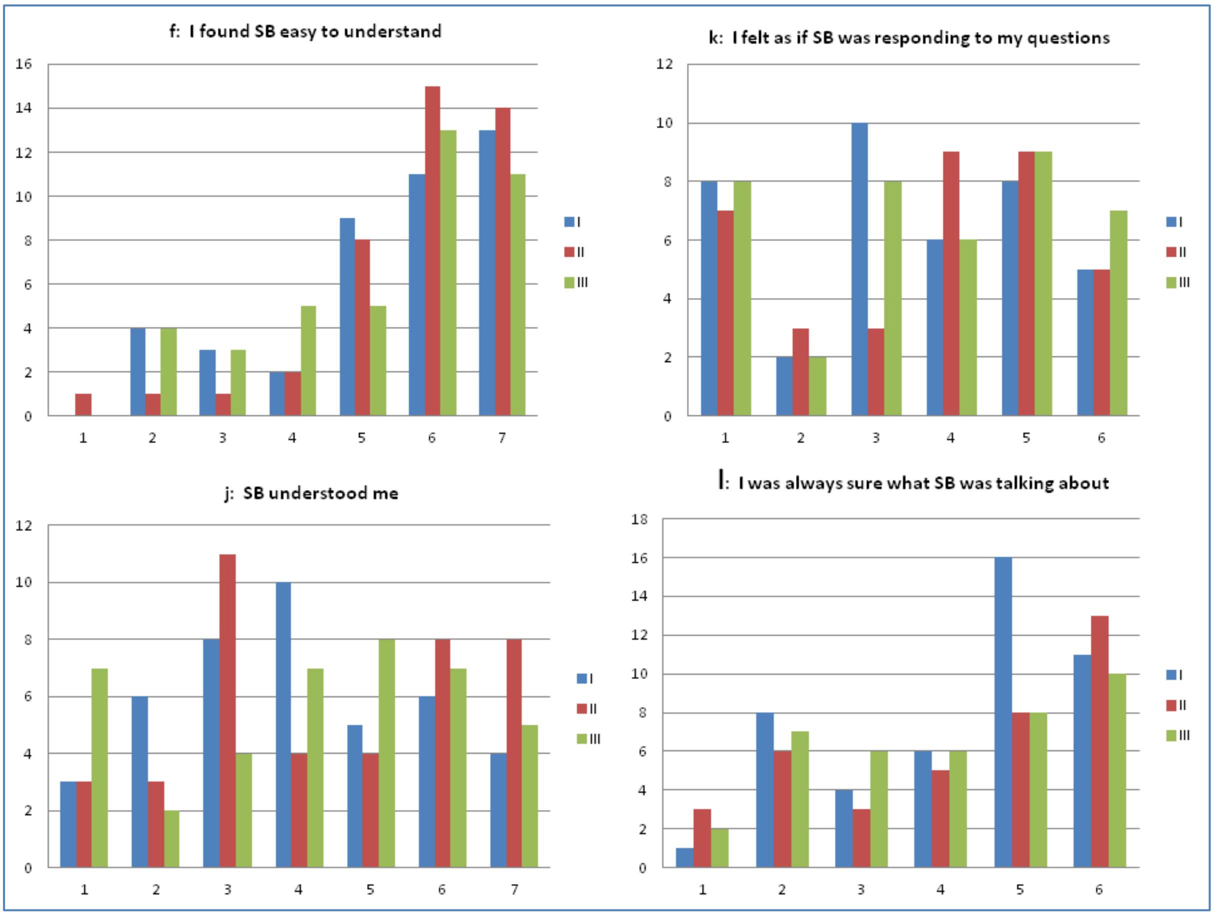


Fig. 3.05. Interaction Ease item response frequency distributions under 3 different versions of the system. Item response categories range from 1 ("strongly disagree") to 7 ("strongly agree") for items **f** and **j**, and from 1 ("strongly disagree") to 6 ("strongly agree") for items **k** and **l**.

Table 3.03 shows that the three systems were perceived to be similar in terms of clarity, but overall, more respondents were happy with the single-threaded System II (88%), compared to the multi-threaded Systems I or III (~79% and 69%, respectively). The user's perception that SpaceBook understood *them* was broadly similar across the three systems, with a much lower rate of satisfaction; only approximately half of respondents agreed with this statement. As for item **f** (I found SpaceBook easy to understand), the single-threaded System II had higher average ratings than Systems I or III. The majority of respondents (almost 2/3) felt that SpaceBook was not responding to their questions, and this response rate was the same, regardless of the System used, suggesting underlying problems with either the ASR or IM, that need addressing. Approximately half of respondents agreed that they were always sure what SpaceBook was talking about, again with System II performing slightly better than Systems I and III.

> "The general quality of understanding was very good"{C14]
>
> The cues were perfectly timed."[B10; task 3; System II]
>
> "info very clear, landmarks and street names helpful."[B09b; task 1; system III]
>
> "'Look ahead of you…' was easy to understand and helpful."[B09a; task 2; System I]
>
> "It was very clear. It told me when it didn't understand me, then I could repeat myself – but only had to do it once."[B06]

> "Good landmarks were provided, that were easy to identify."[B04; task 1; System I]
>
> "It always seemed to know my position and the timing of its updates was good."[B02; task 3; system I]
>
> "directions were reasonably clear with good timing."[B02; task 3; system I]

The manner in which Systems I and III queue up and prioritise different types of information delivery should mean that they are more responsive to the user's location, but any delay in GPS positioning, combined with the queuing effect and any ASR misrecognition problems may mean that the information is not delivered in a time-appropriate fashion, or comes across in jumbled, unordered 'chunks' which confuse the listener, and if persistent, eventually annoy the listener.

> "Delays in positioning affect the experience."[C01]
>
> "Large buildings are good for directions, whereas for smaller landmarks the accuracy of the GPS is critical"
>
> "Poor timing meant that none (of the landmarks) were very helpful." [C02]

When items were analysed separately for association between response and system none of these differences were significant (see Appendix 3.2 for all results).

Some users did comment on the system's use of street names or landmarks to navigate by, with different people having different preferences:

> "It used streets, not landmarks... I preferred street names to landmarks, and I think tourists would, too."[B07]
>
> "Good use of landmarks."[B02]
>
> "It made reference to good landmarks and the directions were quite clear."[B08]
>
> "SB referred to streets I could see the name of, which was enough feedback for me to continue in the right direction."[C04]

SpaceBook's confirmation of progression was valuable to some users:

> "it was helpful when SpaceBook repeated my question. It should then say: 'Is that right?' to verify."[C10]
>
> "simultaneous 'on your right, and on your left' instructions were most helpful."[C08]
>
> "System was confirming things which were easy to see, and was working at the same pace that I was walking."[C06; Sys III]
>
> "voice and information delivery were quite clear"[C05]
>
> " 'Take me to X' seemed to work - it understood a quite human command."[B23]

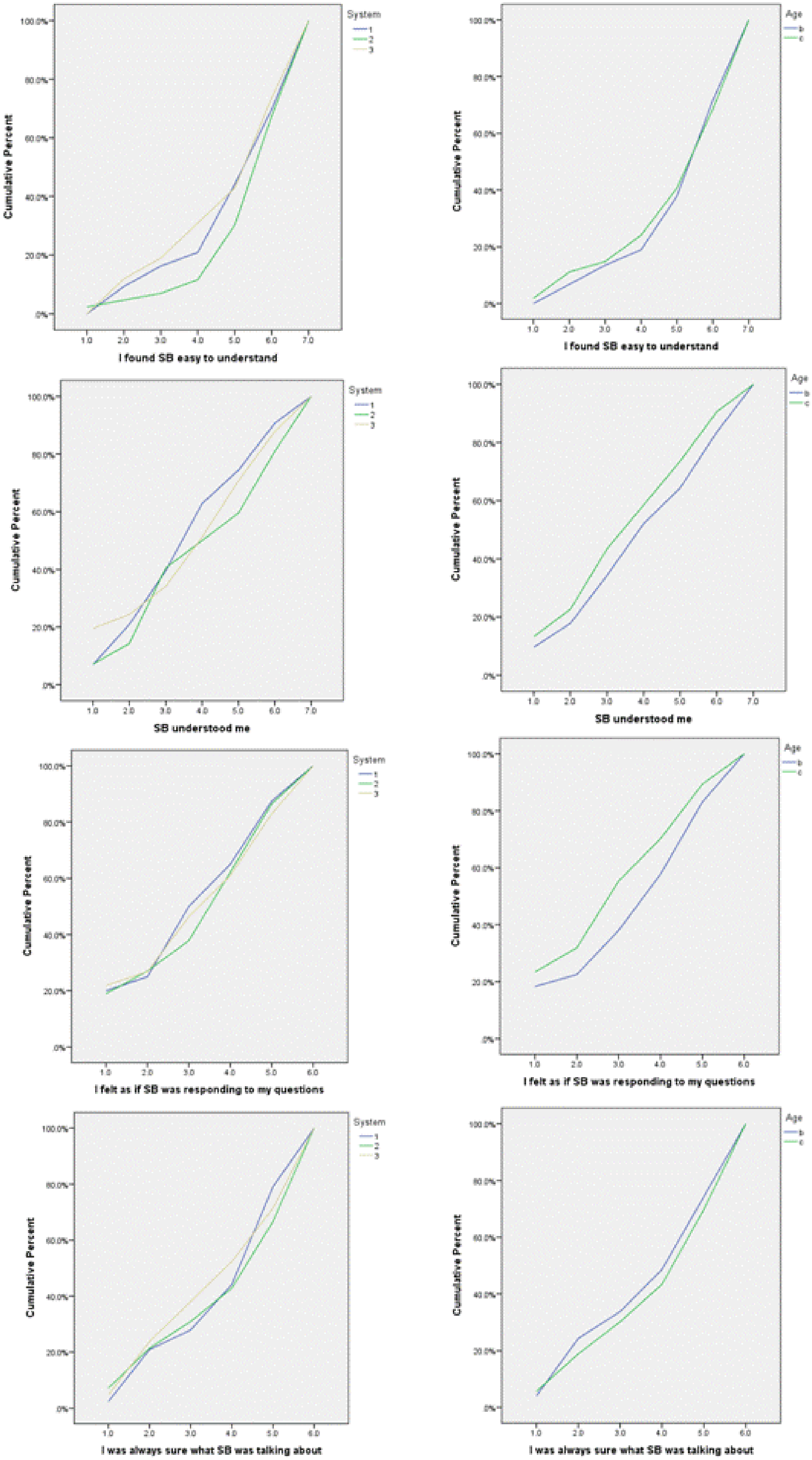


Fig. 3.06. Interaction Ease Scale item response cumulative percent graphs, showing age- and system-effects

Visual examination of cumulative percent distribution graphs (Fig. 3.06) suggests that there is little difference on the basis of System, but that there may be an age effect regarding ease of understanding SpaceBook, or the user's perception that SpaceBook was responding to their questions, with older participants having a lower likelihood of agreeing with these statements than younger ones. However, when analysed using a GLM ordinal regression none of these differences is statistically significant (see Appendix 3.3 for full results).

	System I			System II			System III		
Scale	mean	SD	% 'satisfied'	mean	SD	% 'satisfied'	mean	SD	% 'satisfied'
Interaction Ease Scale	16.98	4.85	69.8	17.51	4.69	83.7	16.47	6.03	65.1

Table 3.04. Interaction Ease scale response descriptive statistics across systems, showing mean, Standard Deviation, and proportion of respondents scoring satisfied for a given scale, where a satisfied score, x_s, is defined:

$$Xs > \frac{scale\ max}{2}$$

These items were combined into the Interaction Ease Scale, and Table 3.04 and Fig 3.07 and 3.08 show the distributions and descriptive statistics under each System.

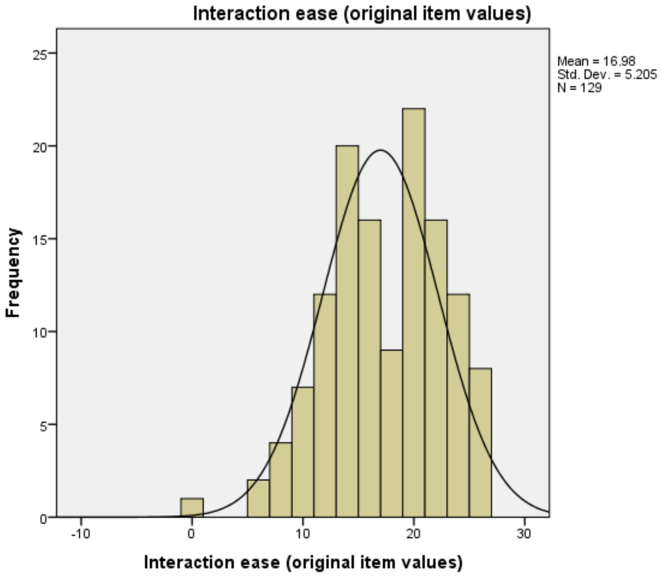


Fig. 3.07. Distribution of total Interaction Ease scale values

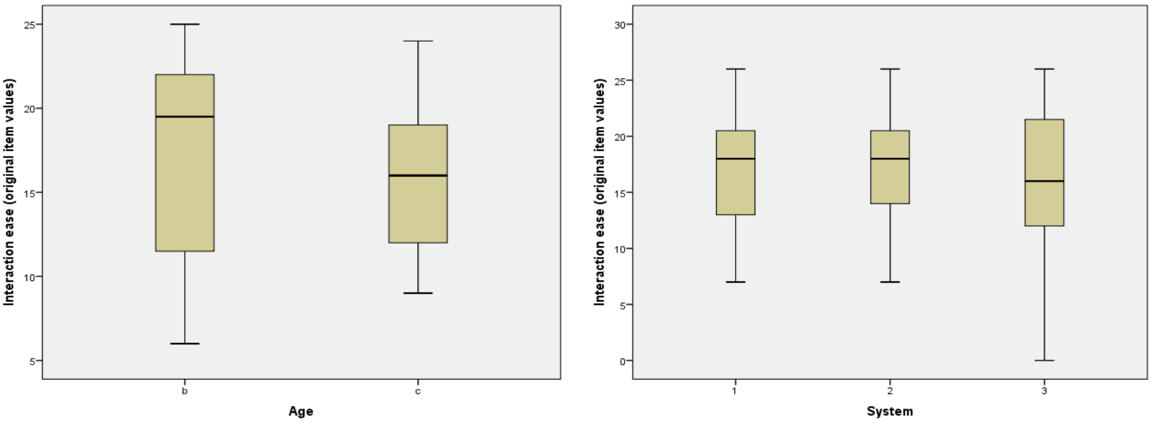


Fig. 3.08. Average confidence and control scale values, for different age groups ('b' = participants aged 18-30; 'c' = participants aged 50yrs+) and System versions

Descriptive statistics suggest an age effect: older respondents tend to have lower Interaction Ease scores than younger respondents (Fig. 3.08), but there appears to be no real difference between System. Analysis of these data through a GLM repeated measures ANOVA showed that neither System ($F_{[2,40]}$=0.479; p=0.623) nor Age ($F_{[1,41]}$=0.703; p=0.407) had any significant effect on Interaction Ease measures, and there was no significant interaction effect between Age and System ($F_{[2,40]}$=0.086; p=0.917).

A general pattern is observable in the satisfaction ratings for the elements of the Interaction Ease scores: System II is preferred by more respondents, with System I and then System III following. One user comments that System II gives the opportunity to speak to the system - an opportunity that is missing in the other systems:

> "On task 1 (system II) it left silences, for me to talk into; on the other two tasks it didn't leave me much chance to speak so I didn't feel as though I had very much input."[C08]

ASR misrecognitions, orientation issues, positioning errors, and lack of specificity in the System's directions all led to negative user experiences, and affected their sense of the system understanding them, or their understanding of the system:

> "...when there were problems because of the way the landmarks were orientated - left and right were sometimes not accurate because I turned around (thus L and R had switched), and they specifically related to my last direction of travel. It didn't correct when wrong. Only references to left and right AT THE SAME TIME help the user to orientate."[C10]

> " Directions which as 'walk ahead' are not useful - you need to know for how far, or until what point is reached.."[C08]

> "SB was not specific in its directions in any of the tasks."[C03]

> "It was as if it hadn't heard me."[C03]

> "Single word commands didn't generate an appropriate response. I was surprised it didn't understand me."[B23]

> "Landmarks suggested by it were not easy to see."[B11; task2; System I]

> "[It didn't understand me]"...especially when starting and I didn't know which phrase to use. 'Can you give me directions to X' was not understood by the system."[B11]

"... I wanted to correct my directions to it and I couldn't repeat them. On my own I would have had to hang up and start again. It kind of ends the whole process when it doesn't understand you - there is no way forward." [B09a]

"It understood location questions, but not direction requests."[B04]

INFORMATION CONTENT

The items related to Measures of Information Content were:

- I found what SpaceBook said interesting
- The landmarks SpaceBook used to direct me were obvious and useful
- SpaceBook gave me the right amount of information

There were two versions of the Information Content Scale: a 2-item version, which contained the first two items listed, and a 4-item version which contained the first two items, and the responses to the last item for both the Explore and the Vista task (*i.e.* twice).

It was anticipated that there would be a difference between Systems I (Visibility Engine on) and III (VE off), as this affects the basic number of things in the environment that the system can tell the user about. The idea is that with the VE on, the system can see as far as the user, and thus can inform them about distant Points of Interest.

Item	**System I**		**System II**		**System III**	
	median	Positive percentage cut-off	median	Positive percentage cut-off	median	Positive percentage cut-off
I found what SpaceBook said interesting	5	57.5	5	52.4	5	61.9
The landmarks SpaceBook used to direct me were obvious and useful	5	53.5	5	52.4	5	54.8
Vista: SpaceBook gave me the right amount of information	3	30.6	-	-	2	30.0
Explore: SpaceBook gave me the right amount of information	2	35.3	-	-	2.5	33.3

Table 3.05. Descriptive statistics of item responses under different systems, showing median item response scores, and positive percentage cut-off values for each item contributing toward the Information Content Scale. Percentage cut-off values indicate the percentage of subjects who rated a given item positively (*i.e.* 5 "somewhat agree" to 7 "strongly agree" on a 7-point scale). All items scored on a 7-point Likert-type scale.

Table 3.05 shows that the interest level of the information content of the SpaceBook system was considered to be higher under System III and lowest under System II, and that there appears to be no real difference in the perception of the suitability of the landmarks used between systems, with only half of respondents agreeing that the landmarks used to direct them were obvious and useful. On the Explore and Vista tasks, neither System I nor System III appeared to have any advantages in terms of delivering the expected level of information to the users, with two thirds of subjects feeling that SpaceBook failed to perform adequately at this task.

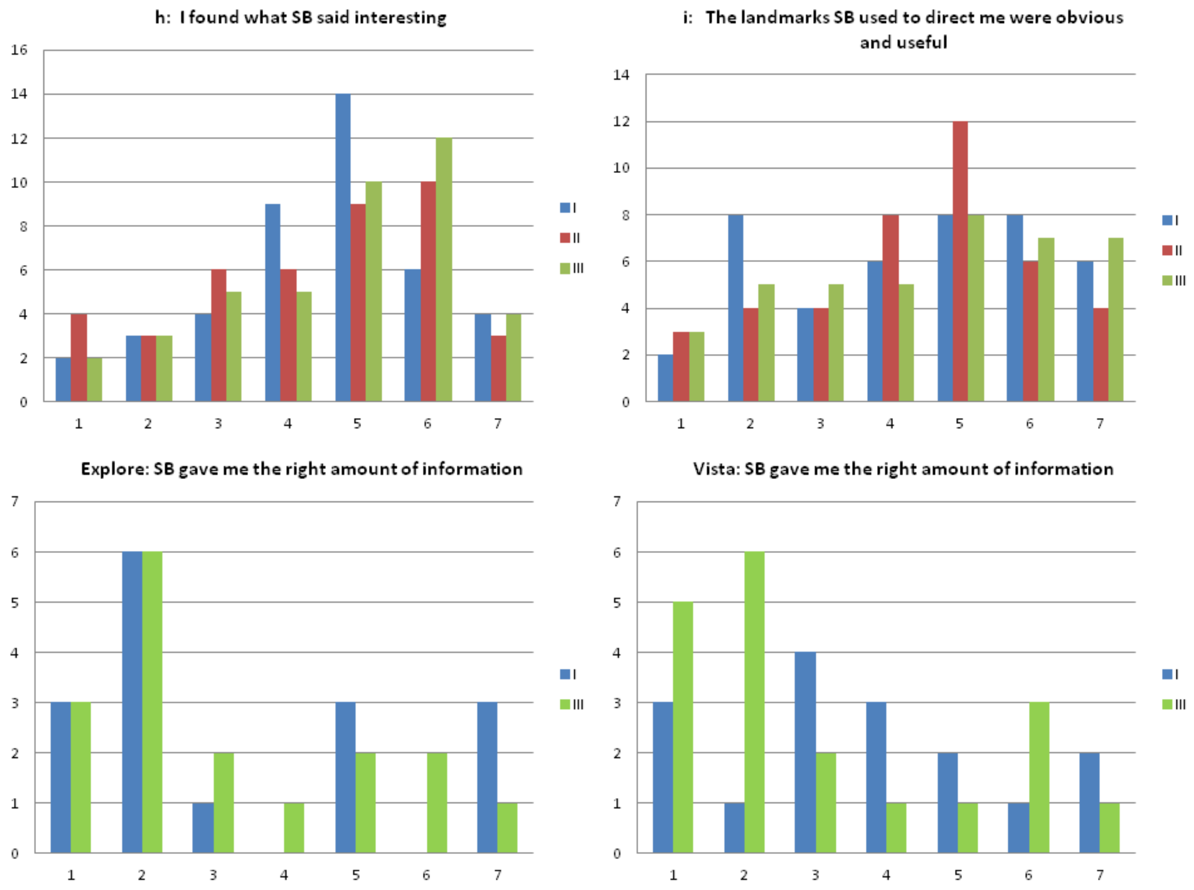


Fig. X. Information Content item response frequency distributions under 3 different versions of the system. Item response categories range from 1 ("strongly disagree") to 7 ("strongly agree").

When items were analysed separately for association between response and system, none of these differences were statistically significant (see appendix 3.2 for full results).

> "There was lots of information about landmarks as we progressed.... Landmarks were described and identified at correct points in my journey."[C12]
>
> [I liked task 2 best]..." because of the extra information en route e.g. Princes St gardens."[B19]
>
> ""Tell me more"[was understood well][B18]
>
> "I liked the details about the buildings we were walking by."[B13; Task 1; System I]
>
> " I was able to ask for information and then further information if I wanted."[B13; task 3; System III]
>
> "It gave me information about the things I asked it to, not about too much other random stuff... instructions were very direct and not overly bogged down with trivia or facts." [B13; task 3; System III]
>
> "SpaceBook was giving information a lot and I could pick out what I needed. It's more comforting to hear a voice all the time than hear silence and wonder what's going on."[B12; System II]

Visual examination of cumulative percent distribution graphs (Fig. 3.10) suggests that there is little difference on the basis of System, but that there may be an age effect regarding the user's perception that SpaceBook was giving them the right amount of information, with older participants having a much lower likelihood of agreeing with these statements than younger ones, and this is confirmed by the ordinal regression analysis, where age is shown to have a significant effect on users' perceptions that SpaceBook was giving them sufficient information for the Explore task ($\chi^2_{[2]}$=14.517; p=0.001), and for the Vista task ($\chi^2_{[2]}$=7.497; p=0.024). See Appendix 3.3 for all results.

There is the possibility that the Visibility Engine creates a certain amount of confusion, particularly if it can see 'too far' - the system then references a Point of Interest that it can 'see', but that is not necessarily visible to the user, or not in the format which the user might expect. For example, referring to St Andrew's square as being visible, when in fact only the tip of the pillar located in the centre of the square can be seen.

> "At Milne's Court it was using landmarks that couldn't be seen."[B02]
>
> "Landmarks suggested by SpaceBook were not easy to see."[B11]
>
> "When it is talking about things you can't see, or identify, it has an unsetttling effect" [C06]

Users commented that they do like to be told about the city and its vistas, especially as Edinburgh has an exceptional set of these.

> "I would like SB to describe the view to me, and not be so focused on small things all the time. Edinburgh has some good vistas!"[B09a]
>
> "If it's not just about getting from A to B, it should encourage you to pause and look around more."[C06]
>
> " If I was a tourist I would expect increased information, and alerts to look at more."[C15]

Visibility Engine attributes may have been impacted by poor ASR performance; in a number of cases, users asked about an entity that was visible to them, but their request was mistranslated, and they reported frustration when the system was then unable to respond to their query.

> "I was asking for explanations but got no response; I was having to repeat myself loudly which was embarrassing and frustrating."[B12]
>
> "Sometimes SpaceBook was mishearing, and sometimes it couldn't understand or answer questions." [C01]
>
> "...It was mishearing questions."[B19]
>
> "It was very difficult to get it to understand me."[C10]

Qualitative user data suggests that issues surrounding Information Content fall into two broad categories: the choice of objects of interest highlighted by the system, and the content of the information delivered. Regarding the system's choice of objects of interest, some users wanted more control, and suggested a user-controlled menu (e.g. arts, history, culture etc) that could be delivered verbally or by button:

> "[I'd like a]...List of menu options: 'you are here' then asks what you want, or lists areas and points of interest around you." [B23]
>
> "[It could have]...Programmes based on arts, history, shopping, culture."[C06]

ASR problems caused frustration when users couldn't get SpaceBook to recognise what they were asking about, particularly if SpaceBook had recently referenced that particular object itself.

> "It was hearing and responding to questions that hadn't been asked."[B12]

> " 'Which road am I on?' and 'Which road do I go to?' were both misheard completely by (SpaceBook)". [B17]

In terms of the content of the information delivered, some users felt that there was not enough,

> "SpaceBook didn't really give me any information of real interest. Any of my requests for more information got no response, or the first line only. It was not expanding. It could be improved on how much information it gives back - it could be prompted to give more." [C05]

> "None of the tasks gave me full answers to my queries. There were big gaps where interesting information could have been given, but wasn't. Nothing about the Black Watch or about David Hume or the view from the Black Watch Monument or Greyfriars Bobby. It was more frustrating than interesting."[C09]

> "I like guided tours by people. This was a bit slow and dysfunctional."[B19]

> "for the St Giles task it highlighted confusing buildings, giving information but not descriptions. ...it was irritating because I didn't know what SB was talking about and couldn't identify the buildings it was talking about."[B16; System II]

> " It would be nice at arrival [at target destination] for SpaceBook to offer to tell you information 'Would you like to hear more…?'."[B09a]

...whilst others felt that there was too much:

> "I thought too many facts were told without asking; I'd prefer it if you had to ask." [B05]

For others, the information was welcome:

> "You can find things on a map, but you don't have the information, so that's nice about it."[B01]

> "There was lots of useful information (**task 1; System 2**) and it was teaching me new things. I got lost on the Black Watch Monument task, and there was not much information on the St Giles task (**System III**)."[B11]

> " I liked that it talked -- there were no big breaks."[B10; task 1; system I]

> "I liked the landmarks. The description of Camera Obscura was useful."[B10]

> "I liked what it said about the library. There was more information and landmarks there (task 1)".[B05]

> "I liked that it gave small facts, and good directions... it was successful and I felt confident."[B04; task 1; System I]

> "I liked particularly information on the landmarks such as Hub and National Library."[B02; task 1; system II]

> "[I most enjoyed using SpaceBook for ...]The first task, because it just told me about more things on the way."[B01; task1; system I]

suggesting that the user's response to this is a personal one, dependent on their own understanding and knowledge of a subject, combined with their level of interest in knowing more. When SpaceBook draws information on an entity from Wikipedia, the perceived relevance is constrained by the order in which sentences have been placed in the Wikipedia entry. This can cause frustration to the user when the first sentences of information delivered relate to the location ("...it's a bit of a redundant statement to tell me where I am." [C15]) and appearance of an entity, when the user already knows what it is and what it looks like. Some users are also sensitive to the source of the information they are hearing:

> "Remove all references to where the information comes from - it's not relevant, and Wikipedia is not always correct" [C15]
>
> "I liked that it gave Wikipedia as the source of its information" [C14]
>
> "It needs more information. Basically it was looking up wikipedia, but it needs to be fine-tuned - there are other sources of information, too." [C16]

Some users were very task-focused, especially on the navigation tasks, and didn't like to be interrupted with extraneous information:

> "It needs to decide what it is. Is it about getting from A to B or more? Is it a SatNav with 'benefits'? If so, it need to tailor the benefits relevant to the user." [C06]
>
> "I'd like more information on the things it chooses to tell you about; provide the user with options to either get somewhere or find out about things on the way - the mode is chosen by the user." [B08]
>
> "[I'd like]...more directions, rather than facts. Make facts come when asked for." [B05]

One possible way of circumventing this is for SpaceBook itself, in its start-up, to verbally define to the user what it is, and how it works. Some users suggested more control over the modes of SpaceBook, allowing the user to prioritise navigation commands over Points of Interest information,. As for Interaction Ease, the Information Content scores are also influenced by any imprecision or delay in positioning, particularly in areas that have a high density of objects of interest.

> "Positioning was sometimes inaccurate, and delays in telling me things, and there were big pauses in between telling me information." [B16]

Types of information that subjects did want, and tried to 'pull' from the system include:

> "When does it open?"
>
> "How much does entry cost?"
>
> "When was it built?"
>
> "What famous pictures are inside it?"
>
> "Who designed it?"

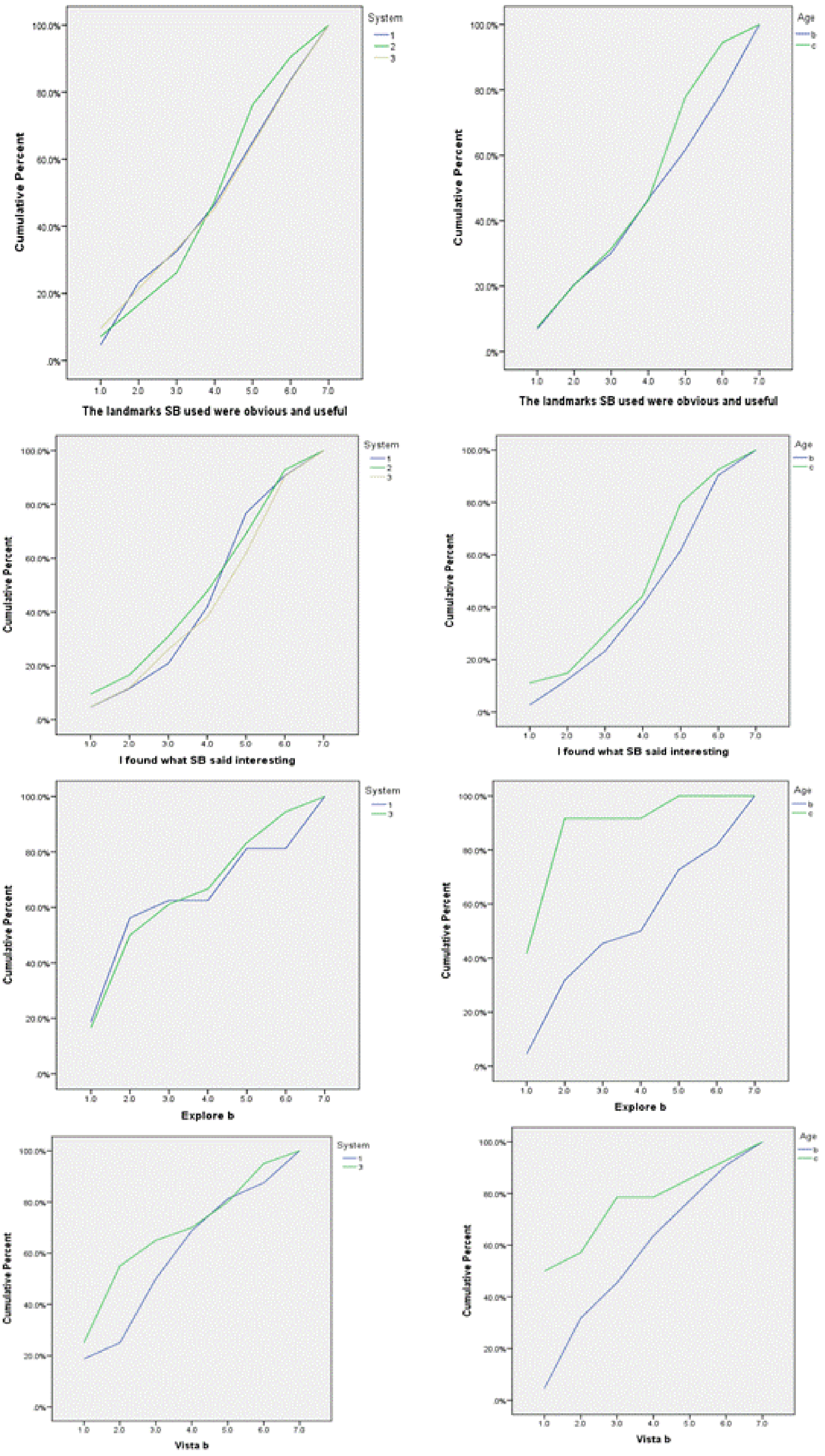


Fig. 3.10. Information Content Scale item response cumulative percent graphs, showing age- and system-effects

These items were combined into the Information Content Scale, and Table 3.06 and Figures 3.11-3.12 show the distributions and descriptive statistics under each System.

	System I			System II			System III		
Scale	mean	SD	% ‘satisfied’	mean	SD	% ‘satisfied’	mean	SD	% ‘satisfied’
Information Content_2	8.98	2.773	69.8	8.49	2.979	67.4	8.88	3.389	67.4
Information Content_4	14.5	5.533	44.4	-	-	-	15.32	5.186	42.1

Table 3.06. Information Content scale response descriptive statistics across systems, showing mean, Standard Deviation, and proportion of respondents scoring satisfied for a given scale, where a satisfied score, x_s, is defined:

$$Xs > \frac{scale\ max}{2}$$

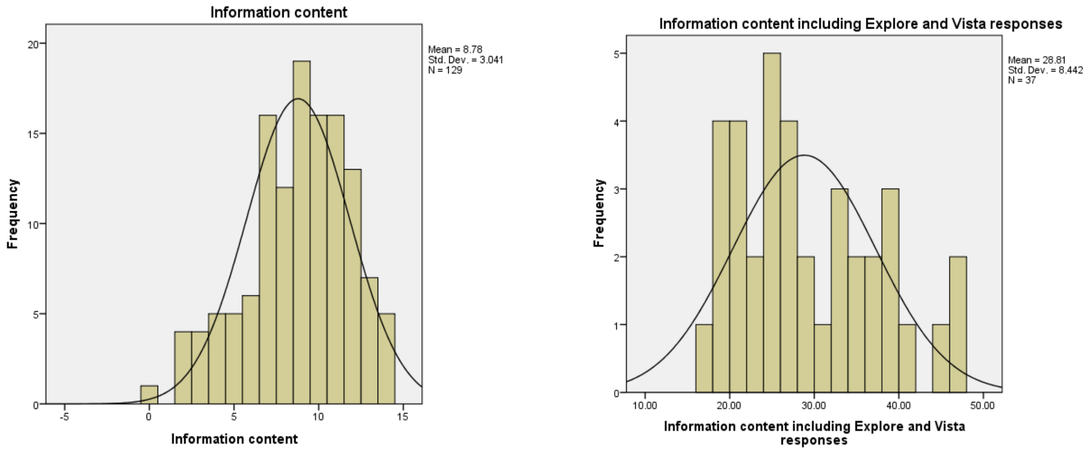


Fig. 3.11. Distribution of total Information Content scale values

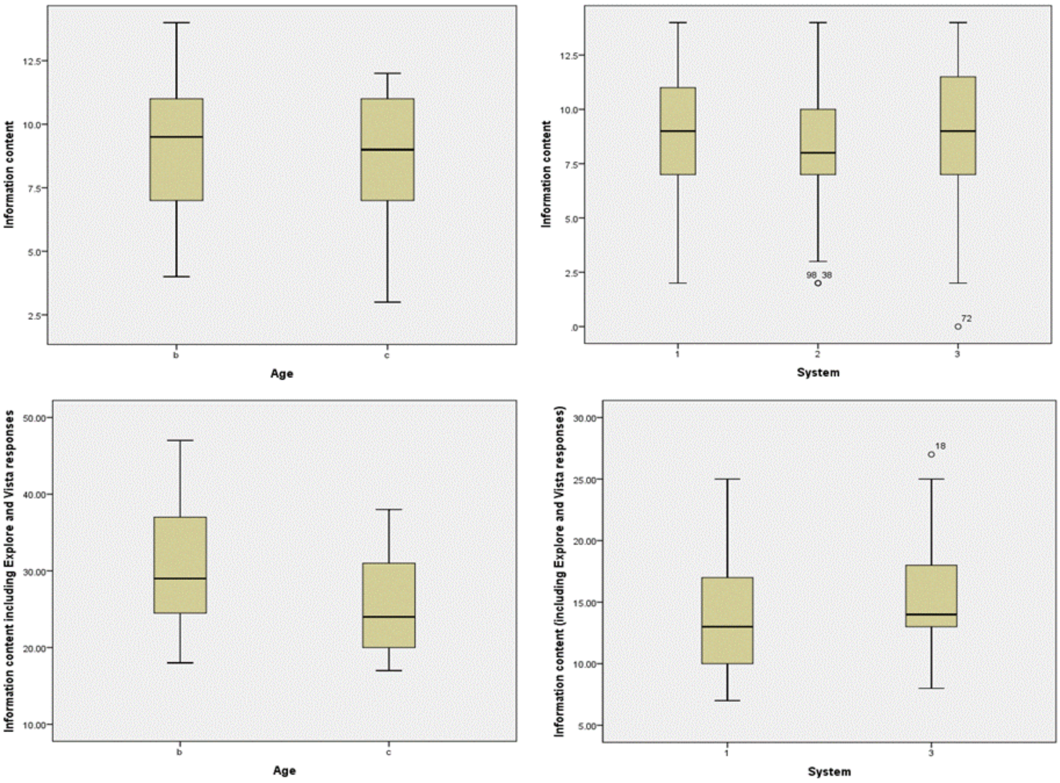


Fig. 3.12. Average Information Content scale values, for different age groups ('b' = participants aged 18-30; 'c' = participants aged 50yrs+) and System versions

Analysis of these data through a GLM repeated measures ANOVA showed that when the questions related purely to the Navigation tasks (*i.e.* the 2-item Information Content Scale) neither System ($F_{[2,40]}$=0.365; p=0.697) nor Age ($F_{[1,41]}$=0.484; p=0.491) had any significant effect on Information Content measures, and there was no significant interaction effect between Age and System ($F_{[2,40]}$=1.776; p=0.182). However, when the respondents were asked to consider information content in the Explore and Vista tasks, as well (*i.e.* the 4-item Information Content Scale), System did not have any significant effect on response ($F_{[1,34]}$=0.00; p=0.989), but Age did ($F_{[1,34]}$=6.495; p=0.016). There was no significant interaction between System and Age ($F_{[1,33]}$=0.390; p=0.537).

3.2 DIRECT SYSTEM COMPARISONS

At the end of the experiment each participant was asked five questions in which they were asked to state their preference for a particular system. Fig 3.13 shows the distribution of those preferences.

When asked to compare directly, there was little difference between Systems I and II on the basis of interest. More respondents thought that System III was the *most* helpful system, and nearly a third of respondents identified System I as being the *least* helpful, although nearly the same proportion of people could not identify which system was the least helpful. The system that most people enjoyed using (38%) was System II, and the system where the greatest proportion of respondents felt most in control was System III. Respondents were the least likely to find System I enjoyable to use, or to feel as though they were in control when using it.

Of the questions listed in figure 3.13, χ^2 analysis of these frequency data, including 'no preference' responses, showed that for the question: 'Which system did you find the most helpful?', there was a significant preference for System III (Table 3.07), and for the question: 'Which system did you find the least helpful?', far more respondents than would be expected were unable to answer this question. This could be due to a number of reasons: reluctance to respond to a negatively-phrased question; politeness; negative overall experience due to a factor other than the features of the systems (e.g. weather, ASR malfunction).

Question	**χ^2 statistic**	**df**
Which system did you find the most helpful?	8.328*	3
With which system did you feel most in control?	2.317	3
Which System did you find most interesting?	1.085	3
Which system did you find most enjoyable?	2.318	3
Which system did you find the least helpful?	13.408**	3

Table 3.07. Frequency analysis results for direct system comparison questions, asked in a debriefing session at the end of the evaluation. * indicates significance at the *p*=0.05 level; ** indicates significance at the *p*=0.01 level.

When only stated preferences are analysed, there is no significant system preference for any of the questions listed in Table X.

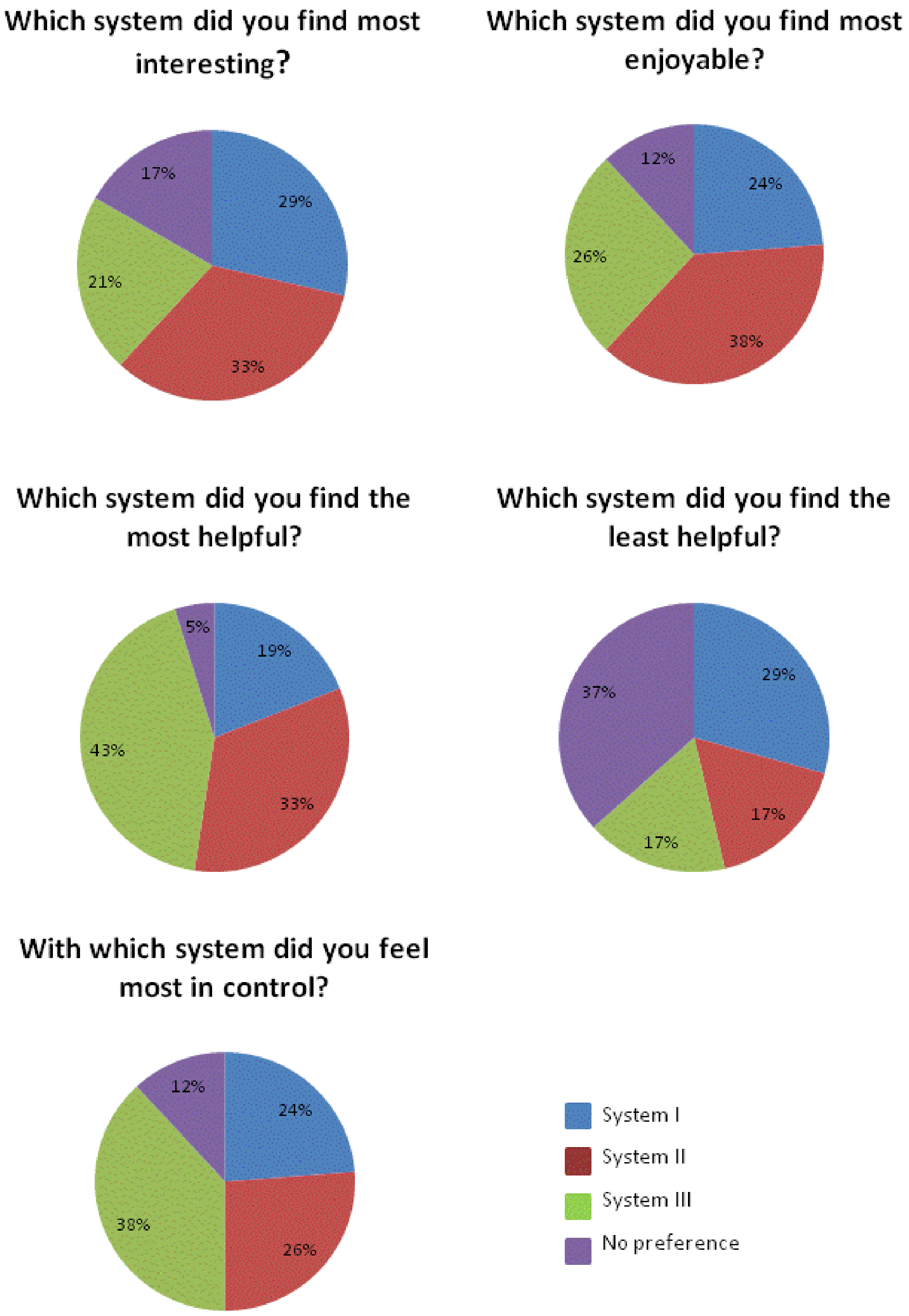


Fig. 3.13. Respondents were asked to indicate direct system preferences for a variety of questions.

3.3 TASK AND LANDSCAPE

It was clear during the experiment, and from the results distribution, that the geography of the task may have an influence on the users' perception of any of the scales listed above. The three tasks could not be identical in terms of points of interest quantitatively or qualitatively, and the topography of the landscape being travelled through may affect the user's sense of confidence or control. Steep, winding routes, where visibility is limited to a smaller radius may mean that the user feels relatively more reliant on the system, and is more sensitive to any reassurance, or lack thereof, on positioning that it gives. Conversely, simple, or open routes, where large vistas are presented to the user, may create issues surrounding ease of interaction, where the user's expectations of information delivery may be higher.

Users themselves commented on the differences between the tasks, and also on the effect that prior experience has on a current task:

> [I found Camera Obscura easiest to find]..." -- maybe because it was a particular route -- straight lines with one turn. The system was clearer."[B14]
>
> " it had proven itself reliable"[B14; task 3]
>
> [I found Camera Obscura easiest to find]..."because of the way SpaceBook gave me landmarks. It's more to do with Edinburgh's City Structure. It's easier to walk down George IV Bridge and see landmarks than around the curved Ramsay Lane."[B09a; task 1; System III]
>
> [*which place did you find easiest to find, and why?*] "St Giles cathedral - it was close and there were no turns."[B03]
>
> "The Black Watch Monument didn't tell me much, it was a very short walk."[B01; task 2; system II]
>
> "It was unnerving at first, but once I had learnt system and taught myself to use it properly, at that point I had more faith in it and felt more in control."[B23]
>
> " I was used to the way it operated."[B20]

The effects of task were investigated through GLM univariate analysis of variance model, with *Age*, *System*, and *Task* as fixed factors, for each of the following scales:

- Confidence and Control
- Information Content
- Interaction Ease

When looking at Confidence and Control, Navigation task 2 had persistently lower mean values than tasks 1 and 3, for all combinations of age and system (Fig. 3.14), and there is a possible interaction effect between task and age under System I, as mean scale responses for task vary between age groups. However, when tested these trends proved not significant (Appendix 3.5).

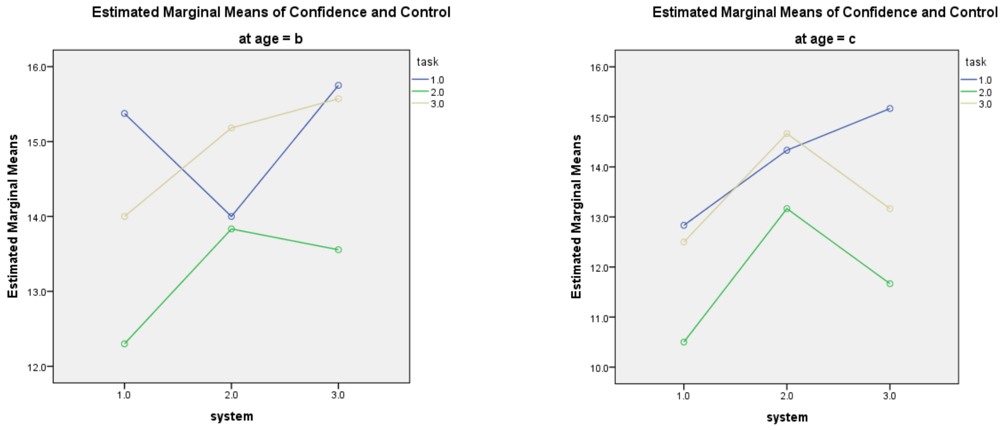


Fig. 3.14. Effect of System and Task on mean Confidence and Control scale values. Key: age b = students; age c = 50yrs+.

When the 2-item Information Content scale was analysed for task effects, there was a marginally significant effect of task on scale values ($F_{[2,110]}$=2.964; p=0.056) and *post hoc* tests showed that there were significant differences between task 1, and tasks 2 and 3. Visual examination of the profile plots (Fig. 3.15) suggests that Task has an effect on the way that younger respondents rate information content when working under System III, returning higher mean values when working on Task 1, and the lowest values when working on Task 2. In contrast, Task has an effect on the way that older respondents rate information content when working under System I (and no effect when working under System III).

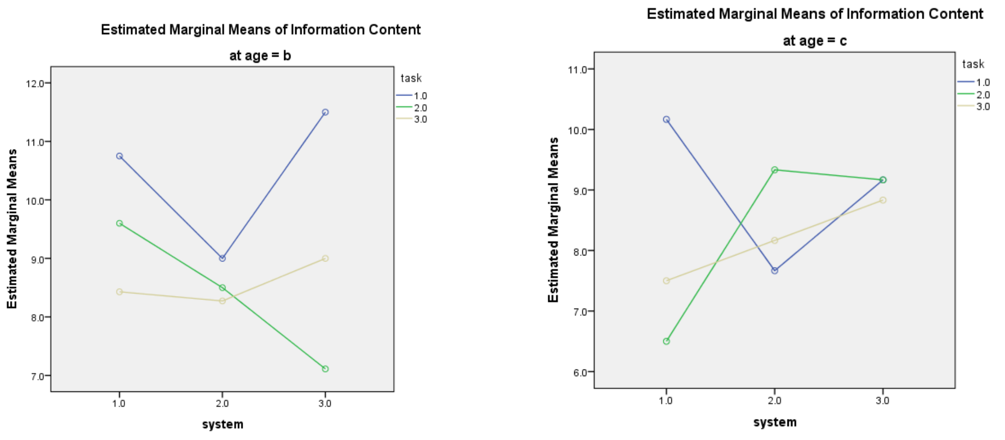


Fig. 3.15. Effect of System and Task on mean Information Content (2-item) scale values. Key: age b = students; age c = 50yrs+.

Consideration of the effect of Task on Interaction Ease scale values showed that younger respondents tended to score higher under Systems I and III, for Tasks 3 and 2, but not for Task 1 (Fig. 3.16). For older respondents, System I scores lower than System II for all tasks, and for System III task 1 returns higher scores than tasks 2 or 3. In fact, there is a significant Task effect: task 1 always scores higher than tasks 2 or 3 ($F_{[2,110]}$=3.050; p=0.051).

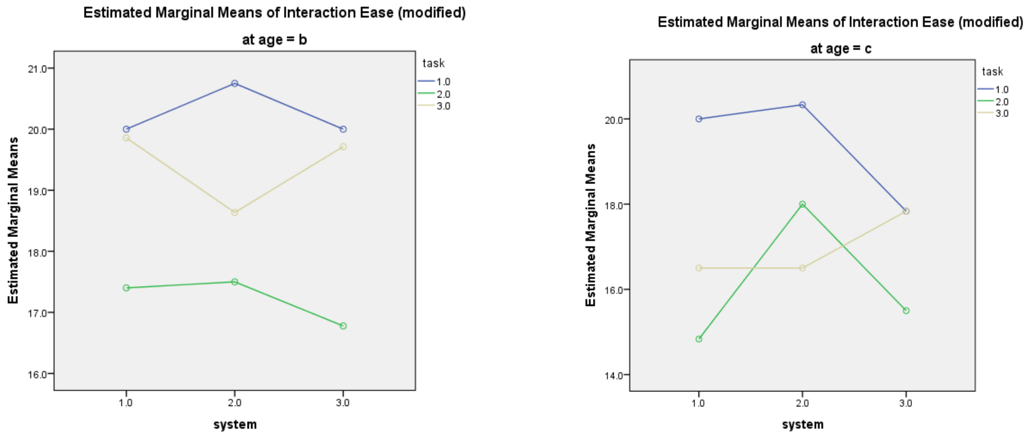


Fig. 3.16. Effect of System and Task on mean Interaction Ease scale values. Key: age b = students; age c = 50yrs+.

3.4 MOVEMENT DATA

Five movement data metrics were analysed by univariate ANOVA for the effects of system and age:

- Average moving speed (m s^{-1})
- Total time spent moving (s)
- Total task duration (s)
- % task duration spent stationary
- Distance travelled (m)

Neither System nor Age had a significant effect on any of the above movement metrics. However, Task did have a significant effect on average moving speed ($F_{[2,84]}$=9.301; p=<0.01) and percentage of task spent stationary ($F_{[2,84]}$=7.983; p=0.001). There was a pattern of higher average moving speed on task 1, compared to tasks 2 and 3, and a greater proportion of time was spent stationary on task 2 than task 1. Although age did not have a significant effect on mean movement metrics, there was an age effect: the main trend that was apparent in these results was that during task 1, older subjects showed greater variation than younger participants, whereas in tasks 2 and 3 this was reversed, with younger participants tending to show greater variation than older participants. See Fig. 3.17 - 3.21.

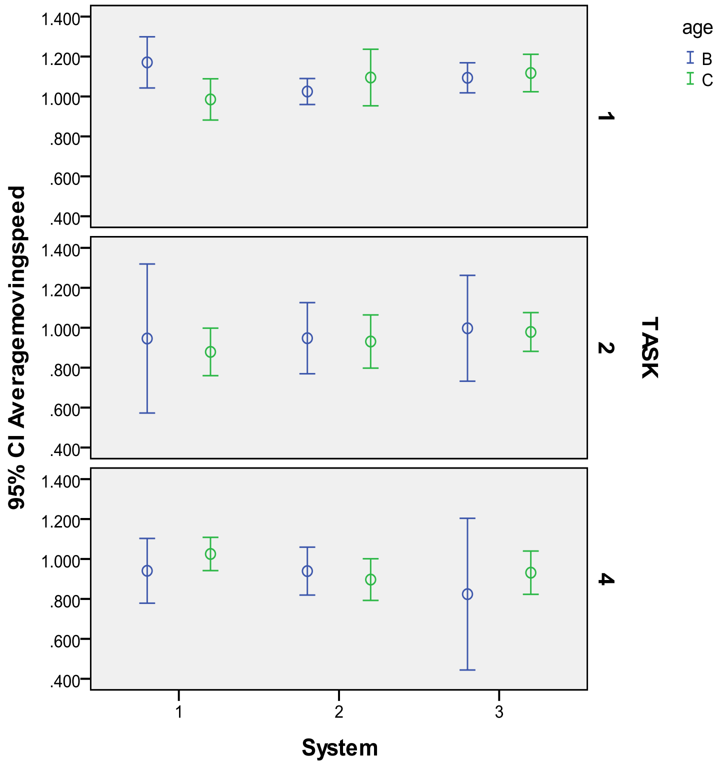


Fig. 3.17. Mean moving speed of subjects under three different SpaceBook systems. B = student participants; C = older participant (50yrs+)

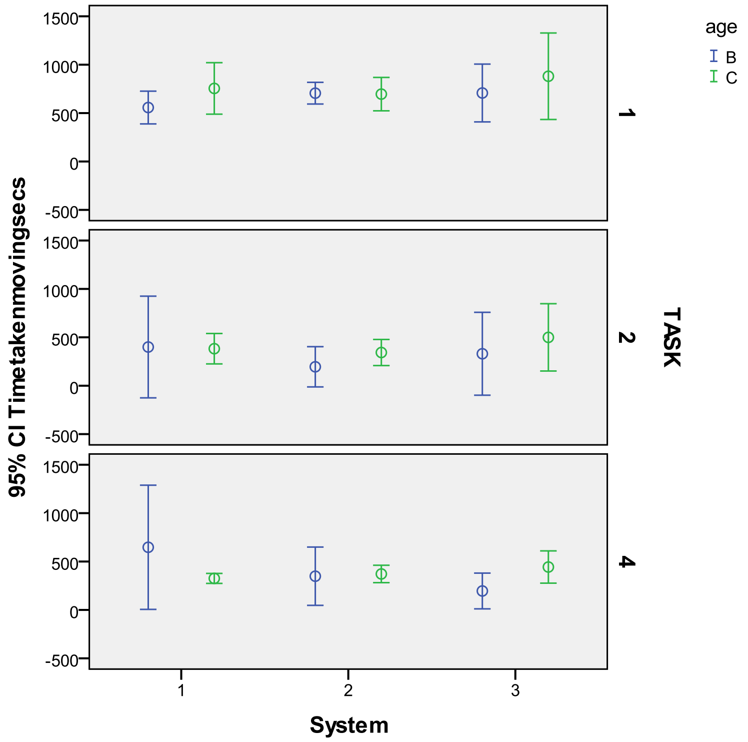


Fig. 3.18. Mean total time spent moving (seconds) of subjects under three different SpaceBook systems. B = student participants; C = older participant (50yrs+)

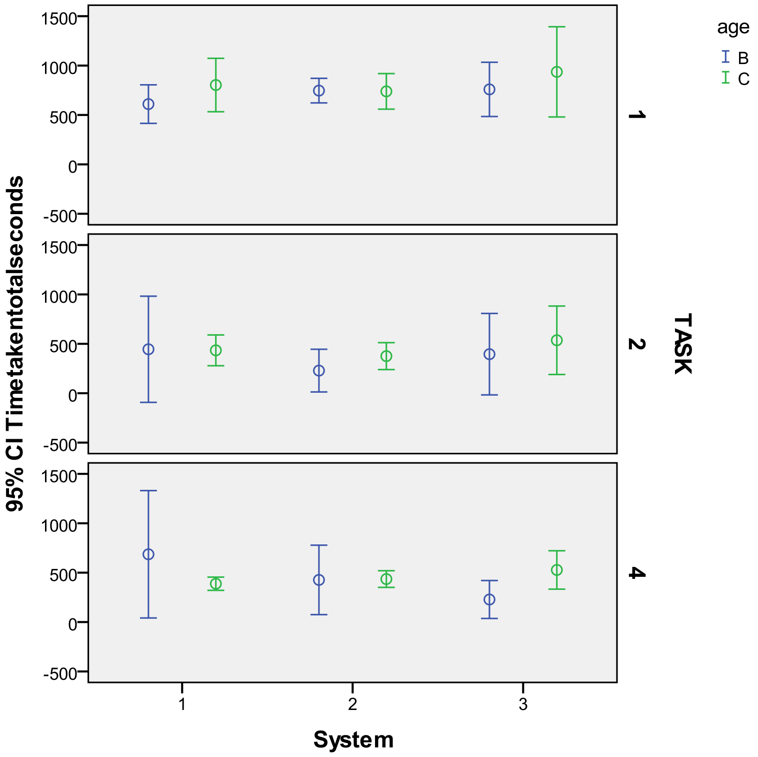


Fig. 3.19. Mean task duration of subjects under three different SpaceBook systems. B = student participants; C = older participant (50yrs+)

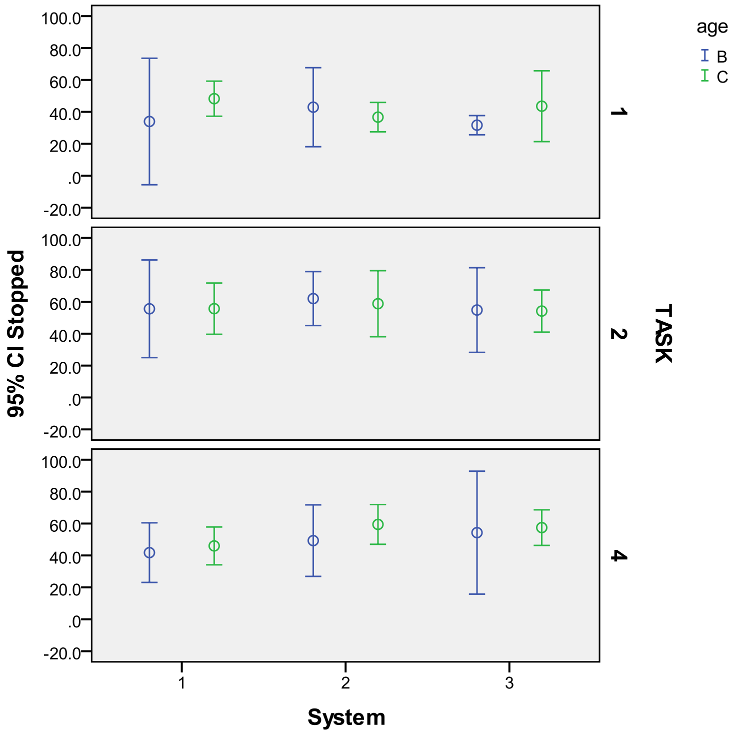


Fig. 3.20. Mean percentage of total task time spent stationary, of subjects under three different SpaceBook systems. B = student participants; C = older participant (50yrs+).

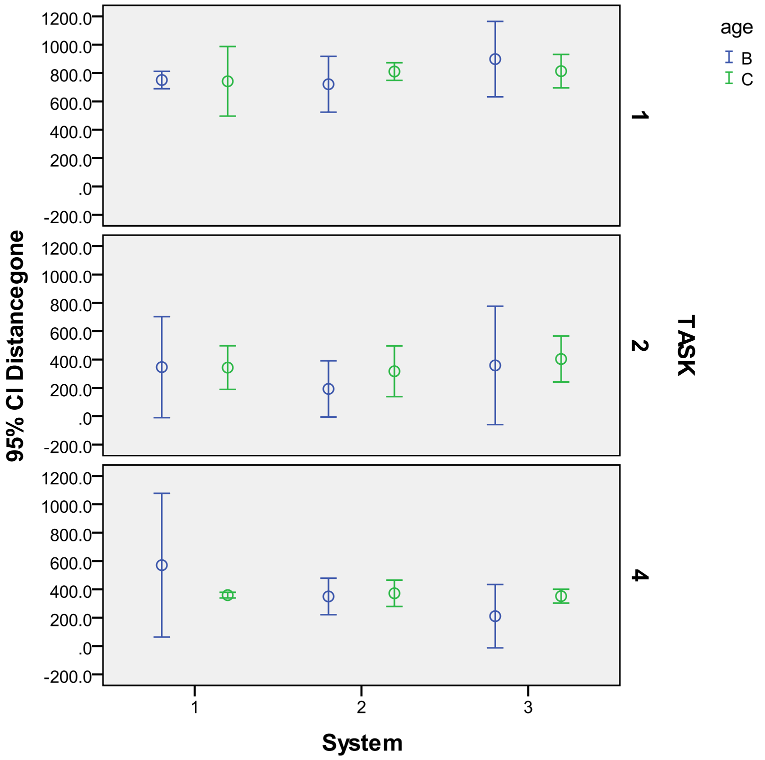


Fig. 3.21. Mean moving speed of subjects under three different SpaceBook systems. B = student participants; C = older participant (50yrs+)

3.5 BEHAVIOURAL CODING

While these did give interesting indications of where, for example, people tended to pause for traffic, or where they might become especially confused, there were a number of difficulties with the approach:

1. When people became lost, they often strayed from the map. Although the maps used contained space around them to address this possibility, it was nonetheless possible to leave the route so that coding on the map was no longer of relevance.
2. When users became confused or frustrated, it was sometimes necessary for the researcher to become involved. This tended to be because of one of two reasons: first, the user themselves would ask a question; second, the system needed to be restarted. Managing user distress and technical issues left little time for accurate coding.
3. The approach assumed that people would follow a generally linear path on the route, moving from the start point to the end point. When a participant became lost, however, and backtracked along the way that they had come, coding on the map could become very chaotic and this could make it difficult to analyse the time sequence of events.

Example: C02; task 1

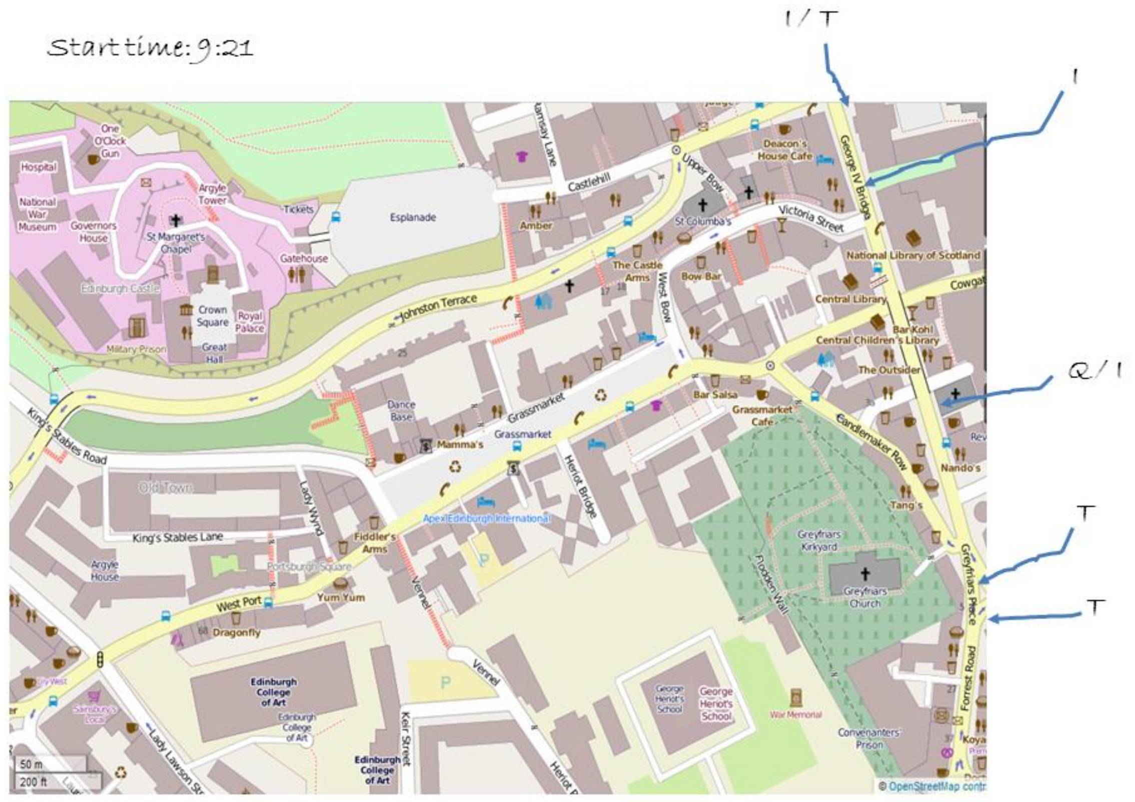


Fig. 3.22 Example of observed behaviour coding, on Navigation task 1

There was generally an inverse relationship, as would be expected, between the coding for confusion and frustration and subsequent user-reported feelings of control (Table 3.08)

Participant	Behaviour observed during task, where: C - confusion; I - Interest; T - Traffic; F - Frustration; Q - Question; LR - Looks around; r/i - researcher intervenes; r/s - system restart			Responses to confidence questions
	Navigation task 1	Navigation task 2	Navigation task 3	
B01	I,T,C	C	T,T	7,7,7
B03	10C, 3F, 4LR	6C, 7F -- several restarts	5F, LR (clear frustration and irritation)	4,2,3
B05	2 T, F	F, 3C	F	6,2,6
B07	c, t	r/s	0	5,7,7
B10	4C	C	0	6,7,7
B13	2 C, r/i*	9C	2 restarts	5,1,7
B14	3C, F, T	C	3C	2,6,5
B15	0	'turn around'	0	6,7,7
B18	7C, r/i, T, LR	2r/s,5C, 2LR	0	2,5,6
C09	r/i, 6C, 5F, 4LR - task failed	0	0	1,1,1

Table 3.08. Behavioural coding results from a sub-sample of participants, and their responses to the post-task question on confidence.

3.6 INPUT PROCESSING: SPEECH RECOGNITION AND UNDERSTANDING

The main finding was that input processing was a crucial but weak component in the process.

Exploration work flow (reminder)

Upon audio input and ASR processing, the recognised user utterance is passed to the parser (*i.e.* spoken language understanding module) and then a semantic representation of the user utterance is passed to the IM. If the utterance is not recognized as pertaining to navigation, the input is sent to the QA server. QA processes the input and sends a response to the IM, for instance a snippet from Wikipedia, a clarification request, or a statement concerning the lack of answer (no answer found, question out of the system's abilities, failure to identify a contextual referent). At any given time, the IM can also pro-actively formulate questions to QA to push information to the user. These requests are processed in a similar manner.

Upon reception of the QA answer, the IM will usually queue the message for generation and text-to-speech output (possibly adapting priorities in the queue order).

INPUT ANALYSIS - EXPLORATION

From audio transcriptions of exploratory tasks, 1192 user pulls were identified, and 452 system pushes. The QA system received 1039 pull requests from the IM and 457 push requests.

Pulls		**Pushes**	
QA system	Transcription	QA system	Transcription
1,039	1,192	457	452

Table 3.09: Identified pushes and pulls

Discrepancies are accounted by the following errors:

- The sound of the user's speech was not picked up by the system (ASR miss)
- The IM handled the input without passing it to QA, usually because ASR mis-recognized the input and the utterance was interpreted as pertaining to navigation. (ASR-IM)
- Noise in the street triggered ASR (ASR false positive)
- In a limited number of cases (18), TTS did not produce the QA answer.

ASR was assessed with a binary marker: either its output corresponded to the user speech or not. In a few cases, slight variations were accepted such as:

USER: *What's near?*
ASR: *what is near me* // accepted

ASR had to be precise enough, independently of the fact that the system could have perhaps rescued the turn despite inaccurate input, as in these examples:

USER: *What's the monument to my right?*

ASR: *what monuments can i see* // not accepted

USER: *There's a large building on my left. Can you tell me what it is?*

ASR: *whats that on my left* // not accepted

However, common errors showed a quite wide discrepancy between user speech and ASR output, such as:

USER: *What is this bronze statue?*
ASR: *what is burns* // not accepted

USER: *No no no, tell me about Jenners.*

ASR: *tell me about gallery* // not accepted

USER: *What's on the hill to my left?*

ASR: *i want a hill* // not accepted

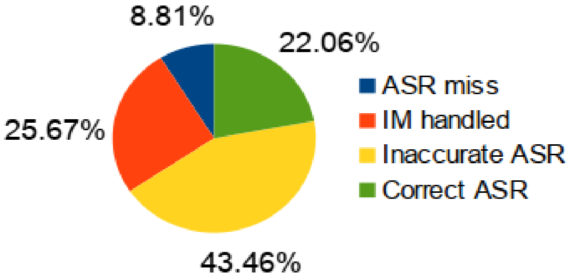


Figure 3.23: Errors types on pulls from transcriptions (N=1,192)

Looking at transcriptions from the user speech (Figure 1), we see that less than a quarter of exploration-related utterances actually reached the QA component in their original form. In 8.81% of the case, the audio was missed by SpaceBook. In 43.46% of the cases, ASR was not correct and in 25.67% the input was handled by the IM but not passed on to QA. The latter seems to also have been caused mostly by ASR misrecognitions transforming the input into non-exploration-related utterances:

USER: *Tell me about whiskey.*	ASR: *Quiet please.*
USER: *La la la la la, tell me about whiskey.*	ASR: *I want a railway station.*).

Now looking at the pulls actually received by QA (Figure 2), less than a quarter of them were actually user utterances correctly recognised by ASR. 7.60% were false positive, triggered by noise or outsiders' voices in the streets. 17.71% were not related to exploration as such but belonged to navigation or general dialogue cues. 94.57% of the latter were actually misrecognitions from ASR:

USER: *I cannot see it yet.*	ASR:*i can see the kirk*;
USER: *There we go.*	ASR: *i don't know.*

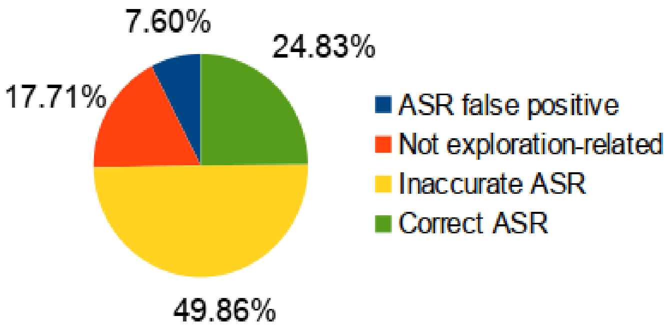


Figure 3.24:Error types on pulls received by QA (N=1,039)

Overall, input processing had a negative impact on exploration as subjects struggled to make themselves understood but not always realising it: an answer could still come out of inaccurate ASR, adding to the confusion.

EXPLORATION UTTERANCES FROM SUBJECTS

1,192 pulls from subjects are available from transcriptions yielding 919 distinct strings. The most frequent were:

tell me more (51)

tell me about the scottish national gallery (14)

tell me about saint giles' cathedral (14)

what is the building on my right? (7)

tell me about princes' street gardens (7)

more information (7)

what is the statue in front of me? (5)

what can I see (5)

tell me about museum on the mound (5)

tell me about jenners (5)

tell me about edinburgh castle (5)

We identified the following content types expressed by subjects' utterances:

1. Explicit questions about points of interest or people, eg.

   Tell me about Saint Giles's cathedral. Who was Adam Brothwell?

2. Requests for more information following a push by the system or an answer to a previous question (*Tell me more.*).

3. Deictic or proximity questions either pointing to a specific point of interest (*What is this statue?*) or asking for for a listing type of behaviour (*What can I see? What attractions can I see? Are there other places of interest?*). Subjects could also be primed by the task brief handed out to them and ask:

   Please find out about three landmarks of interest in this area

   Can you tell me about three objects of interest on the Royal Mile?

4. Questions about instances of a specific type of interest:

A*ny famous artists? Is there a park nearby?*

5. Trivia/factoid questions:

Who is considered Scotland's greatest writer? When was he born?

When was the castle built?

How old is the royal mile?

Why is the mound called the mound?

Those could be of varying complexity and also require coreference resolution:

W*hy is there a heart on the road? Why is the toe gold?* (David Hume statue).

A common pattern was a definition question followed by a factoid.

6. Requests related to exhibitions, access and opening times:

What goes on there?

What happens in the National Library of Scotland?

Can I visit?

When does the gallery open?

7. Clarifications or disambiguating statements, which could be expressed after a question upon the system answer, or before a question:

I can see a big spire. What is that?
It has a little spire with a unicorn on the top of it.
There's a church by, er straight in front of me. can you tell me which church that is?
No not that one. the small one.

Syntactically, subjects generated full sentences but also keyword-based requests, and phrasal coordinations where the question was implicit, such as:

Adam Smith

North Bridge

and the statue in front of Saint Giles?

and the railway station?

We performed a preliminary annotation of user utterances to identify the distribution of co-referring expressions, and how subjects referred to the context and their environment.

We distinguished the following types:

1. **Anaphoric**: a term in the utterance refers to a previous mentioned element.
   When was he born?

   When was it built?
   Are people allowed in?

   What kind of announcements?

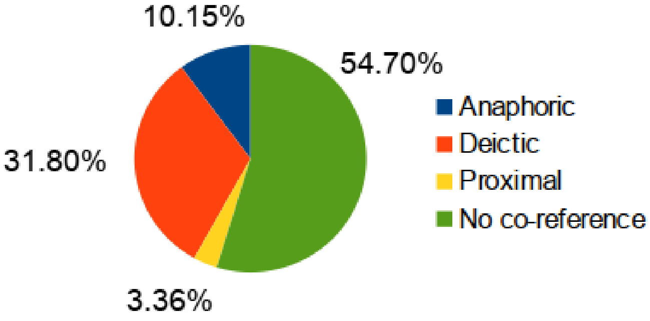
   Are there other places of interest?
   Tell me more. (implicit anaphora)

2. **Deictic**: a term refers to a point of interest in the user view-shed.
   What is on my left?
   What's directly below me?

3. **Proximal**: a term refers to a nearby point of interest, but not necessarily visible.
   What landmarks are nearby?
   Is there anything interesting in this area?

Subjects did struggle with input processing and this affected the way they communicated: they had to repeat, reformulate, explicit and simplify. This is likely to have affected the distribution of explicit versus co-referring utterances. Taking this in consideration, Figure 3.25 shows that contextual input accounted for a bit less than 50% of the input, and when subjects made a contextual reference, it was primarily deictic.

Figure 3.25: Distribution of co-referring utterances (N=1,192)



Subjects otherwise referred to points of interests by their names (e.g. *Tell me about Calton Hill ? What is the Heart of Midlothian?*). This shows at least some familiarity with the city that may not be granted with actual tourists.

INPUT ANALYSIS - NAVIGATION

Problems with speech-recognition and environmental noise for an "always-listening" system meant that the ASR struggled to correctly identify user utterances. This initial barrier to interaction tended to be related to issues such as wind level, traffic noise, the subject's accent and the way in which they phrased their questions, and therefore varied widely between participants.

In order to produce an 'always listening' system (*i.e.* hands-free with no 'push-to-talk' button a grammar-based ASR component had to be deployed, with a specific vocabulary and grammar-based language model (non-grammar based ASR solutions such as Google Speech require a button push and would not be hands-free, and they would also not necessarily recognise unusual place names). This grammar-based ASR was iterated every few days during system development. However, the full range of user utterances encountered during the evaluation were not captured in the ASR grammar and vocabulary, resulting in misrecognition. This is a typical shortcoming of grammar-based ASR, which would be remedied by intensive and ongoing test-and-refine phases in a commercial application deployment.

On the navigation legs, ASR was often more effective than on discovery tasks: B10's first navigation leg, for example, had one ASR misunderstanding and on two occasions the system responded, "You are welcome" when the wind blew, but the subject only made a total of 7 utterances during the leg; C12, by contrast, following the same route, had nine ASR misunderstandings before the destination was established, and a further 23 during the leg; this was from a total of 45 user utterances, meaning 71% were misunderstood.

We took a sample of participants to analyse in detail. Expecting that the ASR would be most accurate towards the end of the study, after all the additions were made to the grammar (see section on additions) and using younger adults since ASR error rates are higher for older adults (Aman et al, 2013), in order to maximise ASR performance. The results (for B17-24) are shown in Table 3.10:

	total	correct	incorrect	% correct
B24	51	16	35	**31.37**
B23	31	15	16	**51.61**
B22	21	11	10	**52.38**
B21	42	12	30	**28.57**
B20	111	59	52	**53.15**
B19	16	7	9	**43.75**
B18	96	18	78	**18.75**
B17	29	10	19	**34.48**

Table 3.10: ASR error rates for a sample of users. [These figures generated from transcripts, contrasting "System recognised" statements contrasted with the preceding few user statements. If a statement had the desired effect, whether or not the system got it 100% right, it is ranked as "correct". This metric is known as "Concept Error Rate" rather than Word-Error Rate.]

Accuracies ranged from approximately 53% for B20 (note that the user learned to use easily-recognised utterances -- ASR recognition was significantly poorer at the outset) to 18.75% for B18. Some of the lower figures can be explained by environmental conditions (e.g. B21 was using the system on a particularly windy day) or users' speech decisions (B18 continually back-channelled, and B24 used a more conversational approach), but the pattern is consistent: the system accurately or usefully identified only around half of user utterances, and often fewer.

Short utterances ("Yes", "No", "Stop") were more successfully recognised than longer ones, and navigation directions ("Take me to Camera Obscura") were often -- although not always -- accurately recognised.

Users used a range of strategies to get the system to direct them to their target. As well as predicted strategies such as "Take me to..."/ "I want to go to...", users stated destinations with no explanatory surrounding speech, as if they were giving a destination to a taxi driver, e.g. "Camera Obscura" (C10: "Camera Obscura, please"). The system could not manage the destination without context.

Others developed non-optimal but often successful strategies such as asking *where* somewhere was and then asking to be taken there. E.g B04 leg2 sys3:

TOURIST Where is the black watch monument?

SYSTEM You will find The Black Watch Monument on Market Street about 220 metres away.

TOURIST Can you take me there?

SYSTEM I am now looking for directions to The Black Watch Monument.

These strategies did not always work. For example, B23, leg 2 sys1:

TOURIST Where's the black watch monument?

SYSTEM You will find The Black Watch Monument on Market Street about 500 metres away.

TOURIST Which direction do I go?

[The system accurately recognised the user's speech but did not answer the question].

Although various non-optimal and unsuccessful strategies were used, most users either initially or eventually adopted successful navigation directions. 88% used "Take me to..."/ "I want to go to...", and 91% of the navigation legs were successfully completed.

However, low rates of recognition and understanding were an initial and significant barrier to use. As B09(a) commented:

> "It kind of ends the whole process when it doesn't understand you - there is no way forward."

3.7 USABILITY

The usability report was drawn from the 170,000-word transcript of the interactions, user qualitative responses and researcher observations. Where necessary (e.g. in examining the effect of interruptions), the audio recordings were also used. Where figures are given, they refer (unless otherwise stated) to analysis of the full transcript.

The evaluations uncovered a range of usability issues which are reported in this section with recommendations for future versions of SpaceBook. In addition, some usability aspects of the 2012 and 2013 systems are contrasted.

In many cases, 2013 showed significant improvements over 2012. One example of this was the ability of the system to correct people who had gone the wrong way:

Need for correction if user going the wrong way

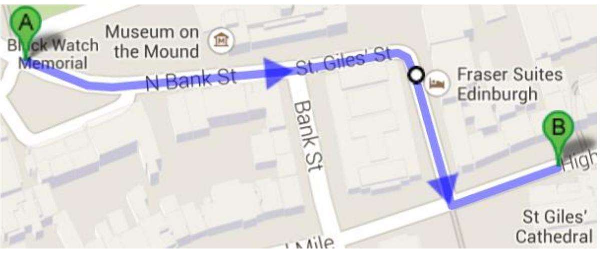
In 2012, eight users (47%) commented on the absence of corrective feedback; if a participant passed their target, there was nothing to redirect them. Participants walked for significant distances (in one case 0.3 mi) before the researcher stopped them.

In the 2013 evaluation, this was far less of a problem. Issues with the system correcting errors was commented on by four participants (B12, C02, C10 and C17), or 9%.

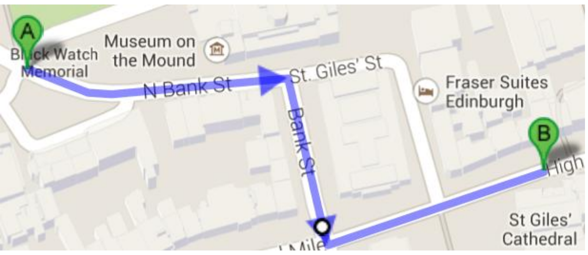
> "If you take a wrong turn, it becomes a problem." (C17).

C02 deliberately tested the system by turning right at the end of George IV Bridge, instead of left as the system had instructed. He had walked to St Giles Cathedral before the system corrected him (approximately 220 ft/ 40 seconds).

Not all the negative feedback was caused by system failure. On C10's Leg 3 (Black Watch Memorial to St Giles Cathedral), instead of directing him to follow Bank Street, the system allowed him to walk up St Giles Street, which also took him to his target destination. He marked the system down although, in fact, it had only demonstrated an ability to update and direct him to the target.



Route taken by user



Route user expected

The junction at the end of George IV Bridge was problematic. Two participants (B07 and B12) continued straight across the junction and down Bank Street (approx. 100 ft) before the system instructed them to turn.

Despite these small delays in correction, the system was far more responsive and effective than the 2012 version.

Only one participant (B08) commented that SpaceBook had failed to inform her when she reached her destination (though in fact the system *had* informed her, but she had not heard).

Clarity of Corrections

However, although the system did correct those who were going in the wrong direction, and attempted to redirect them, this was sometimes unclear. In part, because the system assumed people had obeyed its instruction to turn around.

For example (from the B13 transcript. Leg 2, system 2):

System: **Your destination, the Black Watch Monument, is about 190 metres away, on Market Street.**

> **Continue walking keeping Edinburgh Old Town Weaving Mill on your left and Camera Obscura and World of Illusions on your right. Turn around.**
>
> [user does not turn around]
>
> [System continues as if user *has* turned:] **In front of you, about 10 metres away, you should be able to see the Royal Mile. It is a cobbled street.**
>
> **More on Royal Mile. Edinburgh's most famous and historic thoroughfare, which has formed the heart of the Old Town since medieval times.**
>
> **Continue walking keeping The Scotch Whisky Experience on your right**. [User still has not turned, and is approximately 300 feet away from the Scotch Whisky Experience, walking towards the castle]**. Turn around. You are on Esplanade.**

The short utterance "Turn around" seemed, for several users, to get lost in the surrounding information. Since the system did not confirm whether or not they had turned, and since the landmarks to which it then directed them were *behind* them and so not visible, the instruction was often missed or understood as contradictory.

Additionally, a lack of intonation change and the absence of silent space around the instruction meant that although "Turn around" was a *critical* navigation direction, it was presented like any other piece of information.

The closeness of "turn around" to the subsequent navigation instruction, and the absence of clarifying conjunctions (such as "and" or "and then" or even "once you have turned around") meant that users commonly heard the instructions as contradictory.

e.g. "**Turn around. Carry on down the street**." produced the reaction: "What does it want me to do? Turn around or carry on down the street?"

SpaceBook also sometimes repeated the instruction, which confused users:

> *System* : **Turn around. Turn around. Carry on straight. You will be walking up North Bank Street. Turn around**. (B20 leg 3 sys 1)

<u>Recommendation</u>: The utterance "turn around" should be emphasised to enable users to distinguish it from non-critical information.

RESTATING TASK

An additional issue was that when the system was in the process of guiding the participant, it rarely restated the task when the user required it to. For example, B10 got lost on leg 1 (sys1) and asked: "Where am I?" The system responded: "**You are on St Giles' Street**", but did not restate the target. It then said:

"***Continue walking keeping High Court of the Justiciary on your left, and Coda Music and Scottish Mills on your right. Turn around***."

Had the system explicitly told the user he was heading in the wrong direction, it might have provided a more salient cue to support the direction "turn around".

"STOP"

In 2012 user feedback indicated that a "stop" facility would be appreciated, so that users did not have to listen to a whole segment of information if they were trying to achieve something else, especially if the system had misunderstood them. Commonly in 2012, users would stand waiting for SpaceBook to finish its segment of speech so that they could say what they had intended.

In 2013, the system had a "stop" facility, and it was well-used. The 25 B users used "stop" a total of 181 times. The system recognised it correctly 86 times (47.5%), and failed to recognise it 95 times (52.5%). On 17 occasions the system misheard another utterance as "stop".

The 18 C (older) users used "stop" 78 times, of which only 23 (29.5%) were recognised by the system, and 55 (70.5%) were not. The system also heard "stop" 25 times when it had not been said.

However, these figures for the older participants do not give a fair idea of the system's performance, since some C users repeated the word (particularly as they became frustrated and SpaceBook did not immediately respond),

e.g. C10:

Tourist No, stop stop stop stop. What is- Stop.

System hears: No. Stop.

And C08:

Tourist No, no, stop stop stop stop stop stop.

System hears: No.

If such repetitions are removed, then "stop" was used by C participants 38 times, and the system recognised it 22 times (58%), and failed to recognise it sixteen times (42%).

Some confusions were caused by users putting the command "stop" into a longer context, when frustrated:

e.g. C08: "No. Stop. Stop. Stop. For goodness sake stop it. Enough. Stop."

or conversational, *e.g.*:

B24: "Right okay stop there I -- I see the bank of Scotland. Where do I go now?"

or where the user included the context in the statement, *e.g.:*

B24: "Stop, I don't want to know about the bedlam theatre. I want to get to Saint Giles's cathedral."

B22: "Stop talking. Where is camera obscura?" and B20: "Stop. Where is Saint Giles' cathedral?"

Although "stop" was recognised approximately 50% of the time, it did not always work as the user expected.

"Stop" was understood by the system to mean "Pause while I say something" rather than "Stop talking about that completely", and thus worked only temporarily. After a pause to allow the user to speak, the system re-started the dialogue segment from the beginning and presented it again in full. This could become very intrusive. e.g. from B20's Leg 1 (sys2):

System: **That is the statue of Greyfriars Bobby. It is on your right, about 50 metres away. It–**

Tourist: Stop.

System: **That is the statue of Greyfriars Bobby. It is on your right, about 50 metres away–**

Tourist: Stop.

System: **That is the statue of Greyfriars Bobby. It is on your right, about 50 metres away. It is a statue of a black Skye Terrier dog**.

In addition, "Stop" did not work on navigation instructions, and nor did it stop the *task*, which caused some confusion since most users had not developed a sufficiently detailed mental model of the system to make these distinctions. An example of this occurred for B13 when the system misunderstood her and tried to direct her to Edinburgh Castle:

System: **Okay. I am now looking for directions to Edinburgh Castle**.

Tourist: Stop.

System: **Your destination, Edinburgh Castle, is about 1 kilometre away.**

Tourist: Stop.

System: **Continue walking, keeping Starbucks Coffee Company UK Ltd on your left, and Baked Potato Shop and The Albanach on your right.**

Users were unable to interrupt utterances that the system had already started (e.g. B13 (sys1)):

Tourist: Information about underground vault.

System [at exactly the same time]: **In front of you slightly to your right...**

Tourist: Stop.

System:... **about 40 metres away...**

Tourist: Stop.

System: ...**you should be able to see...**

Tourist: Stop.

System: **Tron kirk...**

Tourist: Stop.

System: **It has a tall spire.**

Although the "stop" command clearly needs fine-tuning to correspond to user expectations and to provide useful functionality such as a permanent or semi-permanent stop on that particular line of conversation, it compares favourably with the 2012 system, in which users could not stop system utterances at all and needed to listen to the full section of information before they could move on.

Recommendation: "Stop" should halt non-critical system utterances entirely.

REASSURING FEEDBACK

One aspect of an audio-only system is the difficulty of providing feedback. Reassurance is difficult to provide because, unlike on a visual map, it is not possible to reassure yourself with a glance; you are dependent on the system to reassure you. Humans during telephone conversations often provide non-lexical, or low meaning, cues that they are still present and nothing is wrong, including 'back channelling' during conversations (e.g. "Uh-huh", "Mm").

There are numerous examples of such back-channelling from the Wizard of Oz experiments that took place in Edinburgh, e.g. from dyad 10:

Tourist Um - I see a Subway

Wizard **Uh-huh**

Tourist I see - uh-huh I see it.

Wizard **Hmm**.

Such supportive cues are difficult for an artificial system to reproduce, and back-channelling is difficult for artificial speech systems to manage.

B18 more than other participants, treated the SpaceBook system like a human and regularly back-channelled small comments, confusing the system, which attempted to treat the comments as meaningful. As a result, the ASR percentage for navigation tasks was unusually low, with only 19% of B18's utterances being understood by the system.

SYSTEM **Turn around.**

TOURIST Turn around, ah.

SYSTEM HEARS Hello.

SYSTEM **You are on Lauriston Place**.

TOURIST Erm.

SYSTEM HEARS No.

TOURIST Find camera obscura.

SYSTEM HEARS Going to camera obscura.

SYSTEM **Carry on straight.**

TOURIST Carry on straight.

SYSTEM HEARS Thanks.

SYSTEM **You are welcome. You are on Lauriston Place. You are on Lauriston Place.**

TOURIST Oh yeah.

TOURIST REC Hello.

Laughter was another non-verbal aspect of communication. Like back-channelling, this had the effect of confusing the system and reducing the success of its ASR.

On 58 occasions, B users laughed. Often (29 occasions) this was when the interaction with SpaceBook was going badly, most commonly because the ASR was struggling but also if the system was insisting on directing the user to the wrong target (5 instances), or had misrecognised "thanks" (3 instances).

C users laughed on 27 occasions; of these, fifteen were due to poor ASR recognition, including two instances of "You are welcome.", and on three occasions users laughed because they couldn't see the landmark referred to, or the system had no information for them (2 instances).

Users often reacted by laughing when the ASR struggled, either because the misunderstanding was funny (e.g. B17: "I don't know where to go." System: **"I heard you say 'I want a beer'"**) or because the user was under stress and failing to get the system to understand what they wanted. Unfortunately, the system was sometimes confused by the laughter and misrecognised it as some other utterance, which only added to the confusion.

Tourist: [laugh] Take me to camera obscura.

System hears: **Repeat that please**.

(B20 leg1 sys1)

Further, humans -- unlike computers -- easily recognise user distress, and know when it is appropriate to introduce new information/ provide reassurance that the user is still being directed to their requested destination; research in this area for stress-recognition by artificial systems remains preliminary.

Thus, providing reassurance is one of the most important, and simultaneously technically challenging, aspect of an audio-only tourism system, and while it was clear that levels of reassurance had improved in the 2013 system compared to 2012, it was also clear that for some users they remained insufficient.

In 2012 lack of reassurance was mentioned by 77% of users in terms of a need to tell the user which direction to walk in, and a need for reassurance that the user was heading in the correct direction, by comparison, in 2013, only four participants (9%) mentioned the need for reassurance in their questionnaire responses:

> "I was at the mercy of this thing that said go left without really explaining. It took… it identified when I first said it your journey is Camera Obscura and then it didn't mention it again until we got up the hill, and I was going, 'are you still taking me to Camera Obscura or are we going to some other [place]?'" (B03 leg 1 sys3)
>
> "Maybe it would be better if I had to ask it to tell me more. I wasn't sure if it wanted me to go down Candlemaker Row -- I needed reassurance I was going the right way, but it was just telling me facts." (B05 leg1 sys2)
>
> C02: "Often it was silent but I'd have liked some reassurance." (debrief)
>
> C08: "SB gave me extra info on where I was so that gave me confidence in the system, but that confidence wavers according to SB performance. For task 3 it got my location wrong. Directions such as 'walk ahead' are not useful - you need to know for how far, or until what point is reached. Need back-up or reassurance from SB that I am on the right track." (debrief)

C08's point is an important one: if users are secure that the system knows where they are, and is directing them to the place they asked for, then these things themselves provide reassurance. When the user is unsure that the system knows where they are, they are more likely to feel insecure. "Walk ahead", with no additional information, is perhaps rather context-free and disempowering.

In general, though, people commented positively on SpaceBook's knowledge of where they were during a navigation task; for example, for leg 1, the most critical in terms of user experience since it was the first, B01, B05, B09a B10, B11, B13, C02, C11, C12, C18, made positive comments such as:

> "It seemed to know where it was straight away." (B01 leg 1 sys1)
>
> "It seemed to know where I was and what I should be looking at." (B10 leg1 sys1)

Improvements to the system and perhaps also the absence of a map-based comparator system, meant that the need for a map was rarely mentioned in 2013: five participants mentioned maps in the en route questionnaires or the debrief (B03, B05, C02, C09 and C10).

Need for reassurance was thus much less of an issue than in 2012, although the transcripts show some participants seeking reassurance from the system during tasks.

On 78 occasions users asked a reassurance-seeking "Where" question, such as "Where am I?" or "Where do I go now?". Eight times, they asked "Which direction?" (e.g. "Which direction do I go in now?") and sometimes to try to get SpaceBook to clarify. E.g. C09:

System: **You will find Camera Obscura and World of Illusions at the junction between Ramsay Lane and Castlehill, about 220 metres away.**

Tourist: In which direction? Where is castle hill?

System: **Sorry. I could not find that in my database.**

Tourist: Er, the camera obscura is 200 metres away in which direction?

Failure to provide reassurance can be critical to task success. B17 (leg1, sys 2) asked the system: "I don't know where to go.", "What am I meant to be doing?", "Where do I go?", and after the system had failed to answer these appeals, the user addressed the researcher:

> "Sorry. I have no idea what, like, it's just telling me where I am and it's not telling me which way to walk... It's just saying, like, this is left, this is on your right and I'm like, no, it's not. I don't, I don't know where to go."

Shortly after that, she abandoned the task.

B14 had a similar period of uncertainty but this was after the system informed him that he'd reached his destination (leg 2, sys 3): The user asked "Where is the destination?"/ "Where is the black watch monument?"

Some participants used strategies for trying to get the system to direct them to their destination, one of which was to request that it directed them to the street. SpaceBook would say: "**Your destination, The Black Watch Monument, is about 250 metres away on Market Street.**" And six of the 43 users asked to be directed to Market Street. (B04, B18, B20, B23, C02, C16).

Several of these users asked multiple times -- B04 asked 7 times in total, and both C16 and B23 asked 5 times. SpaceBook did not recognise "Market Street", and so its responses suggested an ASR error ("**I heard you say: where is mosque. Sorry, I cannot answer this type of question yet.**"), which is perhaps why these users kept asking: they intuitively expected the system to know about something it itself had mentioned.

Asking for directions to Market Street was in itself a sign that other strategies had failed. In the case of B04, using system 3, four separate requests to be taken to the Black Watch Monument ("Where is the Black Watch Monument? Can you take me there?") and a request for directions ("I need directions") had failed to reassure the user that he was going the right way. Because the system was *already* directing B04 to the Black Watch Monument, it simply continued with its directions and did not explicitly recognise that the user required reassurance and a re-stating of the target destination.

Recommendation: A future system should be able to reassure more explicitly, perhaps by restating the question.

CONFUSION OVER WHETHER SYSTEM NAVIGATING:

Associated with the need to restate the task when reassurance is sought, is the need to be explicit about when the system is navigating. The cues to navigation could be subtle (e.g. "**You should be able to see...**") and some participants were uncertain about whether they were being directed or not (e.g. B20 who repeatedly asked for navigation directions even as they were being provided) or that they were *not* being directed; seven navigation tasks (5.5% of the total) were completed without SpaceBook guiding the participant: B23 (leg2), B12 (leg1), B09(a) (leg3), C16 (leg3), C14 (leg3), C10 (leg3), C05 (leg3). Oddly, this occurred most often for Task 3, when one might have expected the user to understand how the system worked. In fact, it is likely that the difficulty of going wrong on Task 3 explains the pattern: users tended to find St Giles without help from the system, and so the lack of directions was not critical.

EXPLORATION/ NAVIGATION BALANCE

In 2012 38.5% of users commented on finding exploration information confusing or distracting when it was embedded in the navigation task.

In 2013, too, some users found the combination of exploration and navigation information confusing, and this sometimes meant they ignored the exploration information (e.g. C13 responding to the question "Did you learn anything interesting?" said: "Not really. I was too busy concentrating on where I was going.")

Users were also still uncertain that the system was continuing to direct them to their destination while it was simultaneously presenting them with exploration information.

> C10 (leg 1 system 1, transcript): "But I want to go to Camera Obscura. Stop diverting me."
>
> "I had no confidence that it was going to take me somewhere so I didn't want to ask it questions in case it distracted it." (C02 -- Leg 3 System 1)
>
> "Awkward. Switched between giving directions and got info instead. It was clear when it WAS information instead of directions but I would have liked to be able to turn it off." (B14 leg1 sys2)

In general, however, users only reacted negatively to the exploration information when it clashed with navigation information they were struggling to hear or understand (e.g. B03 leg 2).

> B20 (leg1, a task which ultimately failed): "It couldn't give me directions because it was telling me about the things we were passing".
>
> B18 (leg1): "I didn't want information -- I wanted to get to Camera Obscura and I was worried about getting there."

So when users got lost and the system continued to provide exploration information, they tended to feel that it should have focused its attention on the critical task -- to navigate them to their destination. In this situation, exploration information, no matter how interesting, became an unwelcome distraction.

An example was B03, who got very lost on Castlehill. The system directed her down Milnes Court, which she couldn't find (it is a narrow doorway leading off the main street into an Edinburgh "close" but B03 was new to the city and unfamiliar with the term). As she got more and more anxious, she found the exploration information very intrusive.

> "It wouldn't listen when I was saying, "Stop" when I was -- and it was like, "and this thing…". It was giving me little bits of history and I was like, "Shut up! I want you to get me to somewhere!" and it would not listen to me requesting it to stop speaking." (B03 -- Leg 2, System 1).

Sometimes the system pushed information that the user regarded as unsuitable, either because it did not contain interesting information or, more normally, because it was inappropriately timed.

Such pushes occurred at various stages; one was at the initial stage where the user was trying to get directions to somewhere:

e.g. (from C02, Task 1, System 2):

> *System*: **I am ready. How can I help you?**
>
> *Tourist*: I would like to visit Camera Obscura.
>
> *System*: **In front of you, about 70 metres away, you should be able to see Bedlam Theatre. It is a church like building with a big red front door.**
>
> **Okay. I am now looking for directions to Camera Obscura and World of Illusions.**

<u>Recommendation</u>: More work needs to be done on balancing exploration information with navigation.

USER CONTROL OVER EXPLORATION PUSH

Some users suggested that exploration and navigation could be balanced for the individual user by giving the user control over when information was pushed.

> B14 "I would have liked to be able to turn it off" (leg1 sys2)
>
> B03 "It was ill-timed, the information it was giving, because I was more interested in the directions. It was fine the first time when I was just walking along but this time just not the time or the place … you need a way to shut off and say 'I don't want to know this information right now.'" (leg2 sys1)
>
> B05 (debrief): "More directions rather than facts. Make facts come when asked for."
>
> B08 (debrief) "provide user with options to EITHER get somewhere, OR find out about things on way - mode chosen by user"
>
> B20 (debrief) "No pushing information"

INTERRUPTIONS

Another source of confusion was incomplete user utterances. People did not speak in perfect English, although they clearly tried to make their speech clearer for the system. However, in addition, timing issues meant that SpaceBook interrupted people while they were speaking, and this had the effect not only of making some participants feel harassed (one of the pilot participants described it as "Like talking to your mum. You know she's going to tell you what she thinks you need to hear whether you want to or not") but it also meant many participants cut short their own utterances so that they were not talking over SpaceBook, and this produced incomplete utterances that the ASR struggled to understand.

> *C12*: The camera ob-
>
> *System hears*: Repeat that.

B07 (leg 2, sys 2) (speech on the same horizontal line is simultaneous)

I need your orientation. Please walk for a few yards and wait until I am ready.

Can you take me to-	**I am ready**
Can you-	**How can I help you**?
Can	
you take me to the Black Watch Monument?	**Behind you, slightly to your left,**
	About ten metres away
(user makes sound of frustration)	
	You should be able to see the Royal Mile. It is a cobbled street.

The system heard B07's incomplete attempts as "Can you take me to church?".

C10 (leg 2 sys 3):

	I am ready.
Uh Black Watch--	**How can I help you?**
Black Watch Monument, please**.**	

System hears: "What's that monument?"

The system's apparent determination to talk over participants could also cause a great deal of frustration. E.g. later in C10's leg 2 (sys3):

I want to go--	**I searched for Walter Montague Douglas.**
(user laughs)	
	Gazetteer for Scotland says Patron of the Arts
I-- I want to go--	**born near Balerno, a daughter**

Just stop. Stop. Stop. **of the Younger family of brewers**

System hears: I want to go to this statue.

Stop **she lived in the Scottish Borders.**

s- black watch, I want to go to the--

System hears: What's that building?

In their seminal 1974 paper, Sacks, Schegloff & Jefferson summarise a basic set of rules for turn taking in conversation. These include the rule that one person talks at a time and that instances of overlap are brief. They note that difficulties with conversational turn taking are common for young children, whereas dominance of the conversation is characteristic of superior social status. The failure to defer to one's conversational partner may therefore imply to users that SpaceBook is socially inadequate, or socially dominant (or rude).

C13 Leg1

System **On the right side of the road, a few metres away, you will see Crag & Tail**.

User I can see--

System **On the left side of the road, a few metres away, you will see Ness.**

User I--

System **Continue walking, keeping Oink on your left...**

<u>Recommendation</u>: Some way of stopping system utterances (perhaps phone-based?) be developed so the system does not talk over the user.

LANDMARKS AND NAVIGATION

SpaceBook depended heavily on landmarks to navigate the user through the urban space. It was therefore critical that these landmarks be easily seen and recognised. In many cases, landmark navigation worked very well (see main report) but there were also usability issues.

Sometimes SpaceBook would use a landmark that was hard to see from the human street-level view. So, for example, a landmark that could only partly be seen: "You said Princess Street gardens -- you can only see a tiny bit of Princess Street gardens over there." (C17 -- Vista -- System 1).

Another example was The Hub, a landmark that could only be seen from one side of the road (commented on by B13, B32, C01, C11) and B01: "... the next landmark after it told me to turn left onto the Royal Mile was The Hub, which I couldn't see yet so that one wasn't very good until I'd walked along, and then I could see it.... There were buildings in the way. I think the road had to curve a little bit. I might have been able to see it from the other side of the road." (B01 leg 1 sys1) . (See fig 3.27 for view)

> "City chambers were not [good landmarks] -- couldn't be seen at time SB pointed them out" (C02 leg2 sys2).

Some participants, who had been on the right side of the road, found the Hub a good landmark:

> "...about half the time they were ok, because they were not always visible - it depends on where the user is standing and looking. Bank of Scotland and the Hub were good landmarks because highly visible" (B08 leg1 sys2)

Landmarks which were named but not described:

> "SpaceBook said to 'keep Writers' Museum on the right-hand side', but you hadn't at that point had any indication where the Writers' Museum was." (C13 leg 3 sys3)

> B13 " The High Court of the Justiciary was not very good." (leg3 sys3).

(B09a, B19 and B24 also commented negatively on the High Court of the Justiciary)

Some landmarks were unhelpful because they were behind the user or visually blocked (B13, C14):

> "some landmarks were visually blocked when I was being told about them" (B02 leg 1 sys2).

At the start of leg 1, the Royal Bank of Scotland was behind the starting point and out of sight (around the corner on Lauriston Road): C08 commented that the RBS "didn't exist" (C08 leg 1 sys2)

> "It was telling me about things behind me." (C02 Leg 3 System 1)

> "It told me to look behind me at the Bank of Scotland -- not very productive." (B14 -- Leg 3 System 1)

> "One [landmark] was around the corner and one was out of sight." (C08 leg 2 sys3)

There was also a problem with some signs not being salient -- e.g. Gladstone's Land (B13 leg 2 sys2), the Hub (C17 suggested it should be described as looking like a church -- leg 2 syst1).

> "Didn't know where they [landmarks] were. The sign for them still couldn't be seen and the Hume Statue was not visible." (C15 leg2 sys1)

One of the landmarks to direct users down Ramsay Lane to the Black Watch Monument, the weaving mill, was not especially salient from a pedestrian's point of view. "I didn't understand what [the weaving] mill was, or where it was." (B05 leg2 sys1). For one thing, "weaving mill" sounded a lot like "woollen mill" (an earlier landmark) and the sign was high on the wall, and not obvious (see black arrow, Fig. 3.26). Google Streetview view of the junction at the top of Ramsay Lane (Camera Obscura is to the right of the picture). The building to the left is the Edinburgh Weaving Mill, which was used as a landmark. However, the sign was not obvious (see black arrow).

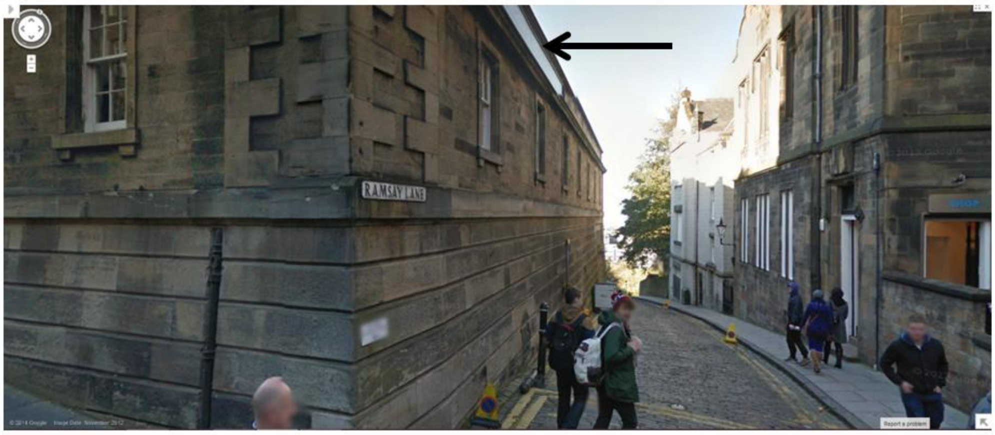


Fig. 3.26: Weaving Mill landmark at the top of Ramsay Lane, used as a navigation point by SpaceBook

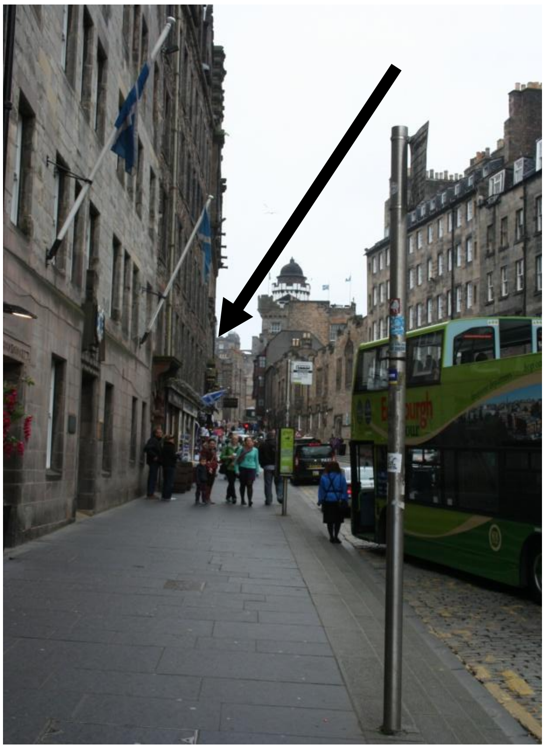


Fig. 3.27 View (or lack thereof) of The Hub from the position at which it is mentioned by SpaceBook as a navigation reference point. Black arrow indicates location of the Hub. See Appendix 2, image (b), for full view

B09a commented: " At start, weaving mill sign is not immediately obvious. When standing at Camera Obscura it's not completely clear which building is the weaving mill" (B09a Leg2 sys1).

> "...it was using the name of some of the woollen mill shops but a lot of them have super-similar names so that was confusing." (B13 Leg2 syst2). C18 also talked about the "woollen mill" when she meant the "weaving mill" (C18 leg 2 sys 2).

Not seeing the mill was also a problem for B21 (leg2 sys1)

Sometimes users' lack of familiarity with Edinburgh terms caused navigation problems. As discussed above, B03 became very lost because she was not familiar with the term "close" to describe a narrow alleyway. [See Appendix 2; image (g) and (j) for examples of Edinburgh Closes].

B07 also commented:

> "It would have been better if it had said Milne's Court was a sidestreet." (B07 leg 2 sys2).

> "It needs to say what Milnes Court is." (C09 leg 2 sys1)

Sometimes because the navigation landmark names were automatically generated from information on the internet, there were unnecessary and potentially confusing additional terms. E.g. Hotel Du Vin Edinburgh (Visit Britain Assessed), Market Cross (Restored), even Camera Obscura and World of Illusions (the additional "World of Illusions" section made the title very long and hard to distinguish)

LANDMARK DESCRIPTIONS

User C17 pointed out that landmarks etc. must be described because visitors to Edinburgh could come from anywhere in the world: "People are not aware that there's such a thing as a 'close' in Edinburgh. You've got people coming off the plane coming from Iraq." [See Appendix 2; image (g) and (j) for examples of Edinburgh Closes].

Where descriptions of landmarks were provided, users responded positively:

> "At first it directed me towards the Bank of Scotland, the one with the green dome. That was quite obvious, I wouldn't have mistaken it – it said 'green dome'." (B01 leg1 sys1)

> "I liked that it used physical descriptions" (B04 leg1 sys1).

> "...when it was talking about the Royal Mile it said it was a cobbled street, and even though most streets here are cobbled it's these small details -- descriptions -- that make you feel you recognise where you are." (B15 Leg2 Sys1)

The advantage of descriptions are also clear from the transcripts:

B10 at the end of leg1 (sys 1) commented to the researcher: "Well, it said [indistinct] white, lighthouse so I'm guessing it's [here]."

B15 at the end of leg 1 similarly commented: "...it brought me here and then it described the building...".

Although it consistently described both Camera Obscura and St Giles Cathedral when participants were close to them, SpaceBook described the Black Watch Monument to only two participants (B14 and B20 -- both in system 3, and only after several exchanges). This was unfortunate as the monument was the least recognisable of the target destinations; users were observed looking around for the statue.

Six of the users (B11, B14, B16, B19, B20, C17) commented on the need for a description of the monument. e.g. "I didn't know what the BWM monument looked like, and SB gave no description, so I didn't know what it was when I got there" (B16).

Users also noted where other landmarks were not adequately described.

> "No description was given of the High Court, and there is no sign at the back of the building" (C15 leg 3 sys2) (See Fig. 3.28, for rear elevation of High Court of Justiciary)

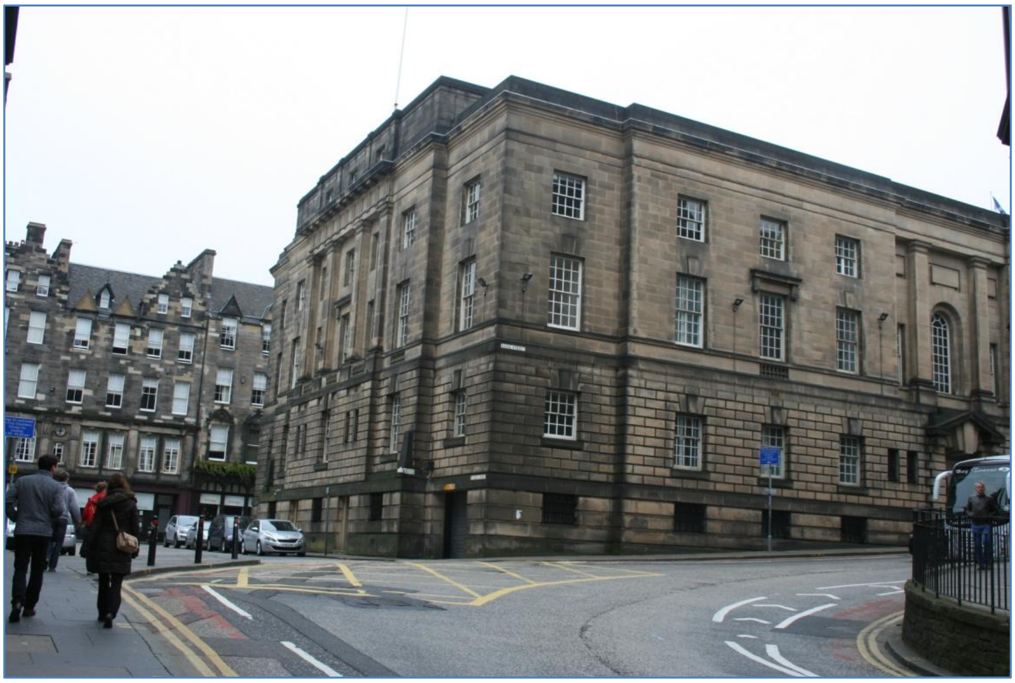


Fig. 3.28 High Court of Justiciary (rear view), as seen by users walking up the Mound on Navigation task 3.

> "Some of them [landmarks] were okay but it didn't describe them clearly enough so I'd know what they were." (B03 leg 2 sys 1)

C02 commented:

> "It's assuming the tourist knows what the landmarks are" (C02 leg 2 sys3)

> "... things like the Adam Smith statue, I had to look quite hard to see what it was. I think coming back to this -- you need in your face names because you've got imagine that you're finding very specific in-your-face names, and you need to say: "Standing outside Deacon Brodie's pub you can see the cathedral..." It used landmarks that you can't assume -- you can't assume people know what things look like." (C17 leg 3 sys3)

> "I couldn't tell how to recognise all the buildings it referred to" (B12 leg3 sys1)

<u>Recommendation</u>: consistently provide landmark descriptions.

LEFTS AND RIGHTS

Ten users (23%) reported that SpaceBook sometimes got their orientation wrong (B05, B11, B13, B15, B21, B22, C07, C08, C09, C16).

> e.g. B13 (sys 1 leg 1): "...it told me "destination on your left" and, like, I could see it on the right as we were coming up the street."
>
> B15 (sys 1 leg 2) "...it told me that I should have a shop which was to my right, on my left"
>
> B21 (sys 2 leg 3): "Cathedral which SpaceBook said was on left, should have been right."

Indeed, the system did struggle with orientation, e.g. from C13 transcript Leg 1 (sys 1):

> *System* **Continue walking keeping Oink on your left, and Howies Restaurant on your right.**
>
> **Continue walking keeping Howies Restaurant on your left, and Oink on your right**.

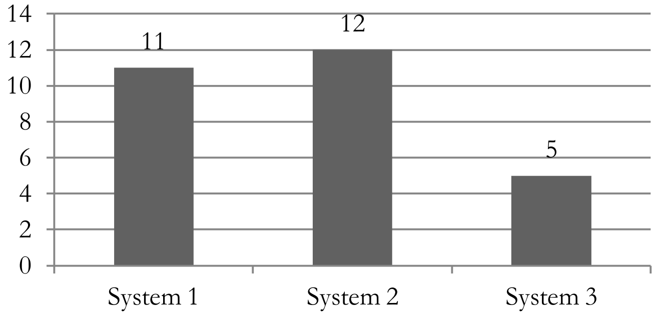
DELAYED CUES

Timing remains an issue for SpaceBook. In 2012, 13 participants, or 76%, commented on cue slowness. In 2013, 19 participants (44%) commented on it. For example:

> "SpaceBook always thought I was behind where I actually was." (B08 Leg 1 Sys2)
>
> "... it was timed slightly off, so it would say there was such a building ahead, 10 metres on my right when it would really be on my right about 10 metres behind me." (B01 Leg 1 Sys1)
>
> "It didn't seem to know where I was. Told me the Bank of Scotland was in front of me when it was behind." (C18 sys1)

User comments on six questions were taken from two questions from each post task questionnaire: speech timing and usefulness of landmarks. Some users commented that more than one system had been slow.

Fig. 3.28 Comments on cue slowness



(B02 commented that all systems were slow)

REPETITION

Users in 2012 felt that SpaceBook did not repeat information often enough; in 2013, the repeat instruction was significantly more sensitive and there were far fewer problems with directions being lost in traffic noise compared to 2012. SpaceBook 2013 was much better at repeating itself. However, users (B03, B06, B09a, B10, B11, B22, B24, C08, C09, C13, C17) in 2013 complained about excessive repetition.

> B10: It said, about five times, "In front of you is the Bank of Scotland -- you should see a green dome" and then it said it again and then it said something like "there is no more information". Something like that, which I wouldn't have minded once, maybe twice, because I was walking and looking at it for quite a while but, yeah, there was a point where I was just like, "Yeah. I know."
>
> B03: "I liked the little, like, bits of history but it would repeat them. So I'd walk past a statue and it would say, "This is blahblahblah" and I'd walk back past the statue and it'd say "this is blahblahblahblah". It's okay -- I've got that information. You can shut up now." (sys3 leg1)

To some extent, repetition was a result of overlapping use of landmarks when landmarks used as a navigational cues were also landmarks of interest. This could produce information that appeared repetitive:

> "**In front of you, about 130 metres away, you should be able to see the National Library of Scotland. It is opposite to Central Library.**
>
> **More on the National Library of Scotland. Located on George IV Bridge in Edinburgh is the main building of the National Library of Scotland, Scotland's only copyright deposit library, which means that a copy of every book published in Britain must be lodged there.**
>
> **On the right side of the road, about 20 metres away, you will see the National Library of Scotland.**" (C02 Leg 1, System 2)

In addition to this kind of repetition was the system's habit of repeating itself directly, e.g.:

> "**You are on George IV Bridge. You are on George IV Bridge**." (C02 Leg1 sys2)

Much of the repetition, however, was attributable to ASR issues. Of 178 instances through both B and C transcripts, where the system repeated itself ("**I repeat...**"), 170 (96%) were the result of ASR misrecognitions.

"YOU ARE WELCOME"

For the B group (B01-24), there were 91 separate instances of the System saying "**You are welcome**". 40 were misunderstandings of something else the user had said, and 25 were environmental noise, most often the wind.

In fact, users very rarely thanked the system. There were nine instances of a user saying "thanks" or "thank you" during the interaction; three of these were recognised and responded to by the system, six were misunderstood.

Even when the system correctly identified the words, "**You are welcome**" was not always a wholly appropriate response.

e.g. (B05 sys 3 leg 3):

> *Tourist* Tell me more about museum on the mound.
>
> *System* **I searched for museum on the mound but could not find any relevant information.**
>
> *Tourist* Thank you anyway.
>
> *System* **You are welcome.**

On six occasions, users said "thanks" or "thank you" and the system misunderstood their utterance and failed to respond appropriately.

For C01-18, there were 34 instances of the system saying "**You are welcome**", of these 21 were misrecognitions of something the user had said, and six were environmental noise (5 were the wind, and one a hammer blow). On four occasions, the system correctly heard the user say "thanks" or "thank you" and responded appropriately. The remaining three instances have no clear cause.

Hearing environmental noise as "thanks", and responding, was especially confusing for participants, who had naturally filtered out noises like the wind blowing. They tended to understand it as the system randomly dropping in "**You are welcome**", e.g. B19 and B13:

> "SB puts in random 'You're Welcome' s!" (b19 leg3 sys3)
>
> "It also randomly just said "You're welcome." B13 (leg 1 sys1).

B05 commented that the "**You are welcome**" made him feel as if he was interacting incorrectly: (leg2sys1):

> "It said, "You're welcome" and it occurred to me I should thank it more."

Finally, B03, who had a very unsuccessful interaction with SpaceBook on leg 2, found the "**You are welcome**"s annoying: (leg3sys2):

> " It says, "You're welcome" in an obsequious manner, like I've said 'thank you' for taking me in the wrong direction."

<u>Recommendation</u>: consider removing "You are welcome" until speech recognition improves.

4.0 KEY FINDINGS AND RECOMMENDATIONS

4.1 KEY FINDINGS

The results presented above suggest the following:

1. Users have a greater sense of control and confidence when using Systems II or III, and less when using System I, regardless of age

2. There is a tendency for users to find interacting with the System easier under System II than I or III, and although there is no significant age effect the Task in question is important: users found interacting with the system easier during Task 1, suggesting geography and distance have an influence.

3. In terms of the information content delivered to users whilst completing the navigation tasks, users did not discriminate between the systems. However, there was a significant age effect, with older respondents feeling that not enough information is being given in the Explore tasks. There is also a task effect - younger respondents found task 1 more interesting than tasks 2 or 3, but only when working under System I.

4. When asked to consider 'helpfulness' of the system, System III was the preferred system, but users did not find any difference between systems in terms of enjoyability.

5. The nature of the task being asked of the user and system is critical to the user's perception of the system. Task 1, which was much longer than the other two tasks, showed higher user moving speeds, users rated the information content higher, and they found it easier to interact with the system. A number of users commented that there were more objects of interest on this task, and that there was more time to appreciate them. Users spent more time stationary on task 2, and this may be in part due to the rather difficult start, where large numbers of users found it difficult to locate the starting turn, and additionally SpaceBook found it very difficult to locate the user (narrow, enclosed streets).

6. Independent of System version, the input processing components of SpaceBook are crucial but weak, and inaccuracies generated in this module have a negative impact on exploration, and users' perceptions of the capabilities of the system as a whole.

4.2 RECOMMENDATIONS

Based on usability, a future system should incorporate the following:

- **"Stop" should halt non-critical system utterances entirely.**
- **The utterance "turn around" should be emphasised to enable users to distinguish it from non-critical information.**
- **A future system should be able to reassure more explicitly, perhaps by restating the question.**
- **More work needs to be done on balancing exploration information with navigation.**
- **Some way of stopping system utterances (perhaps phone-based?) should be developed so that the system does not talk over the user.**
- **System should consistently provide landmark descriptions*.**
- **Consider removing "You are welcome" until speech recognition improves.**
- **Evaluation tasks need to be more directly comparable in terms of length, and density of objects of interest**
- **An initialisation speech from the system, reminding users what it does and how to interact with it would be helpful in managing user expectations.**

5.0 CONCLUSIONS

The purpose of this evaluation was to assess the functionality of particular modules within the SpaceBook System, and to assess the system as a whole from a user's perspective. The evaluation process was informative as it generated very specific user feedback, and highlighted areas of the system where further development is a possibility, shedding light on both user expectations and behaviours, and system strengths and vulnerabilities. The use of subjects who were not familiar with the development of SpaceBook, and who were drawn from a range of backgrounds and ages provided a representative sample of potential users, and illustrated some of the specific issues that may arise from the natural variation in future users' expectations and interaction styles.

The evaluation concentrated on a set of priorities that it was anticipated would be important to users, in order to discriminate between the functionality of particular modules within the system: the ease with which the user can interact with the system, the user's sense of confidence and control when using the system, and the user's satisfaction with the information delivered.

It was anticipated that the 'full'system (*i.e.* System I, with a functioning Visibility Engine, and multi-threaded dialogue) would be preferred over versions where either of these modules were not incorporated, as it allows more control and has more information to access. As it turned out, System I was not preferred in the navigation tasks, and this is possibly due to user desires and expectations. Although System I could 'see' distant points of interest, frequently the user couldn't (high buildings, narrow streets) and so tended not to ask about them (pulls), or got frustrated when information on them was pushed, particularly when focused on reaching the destination. Many users seemed quite task-focused during the navigation tasks, and did not make many queries regarding distant landmarks, being generally more content for the system to tell them information on Points of Interest as they were passed. Thus, for a user trying to get from A to B, information about distant landmarks may be of a much lower priority, and of higher priority is more in-depth information on the landmarks at close proximity. Where input processing problems made accessing this further information difficult, or when the information retrieved was perceived to be of shallow depth, or relevance to the user's interest, users reacted negatively. Users felt that they had less sense of control under System I, possibly because there was a lot of information being pushed at them, and the interaction between input processing failure and the volume of pushes created a sense of confusion. On navigation tasks a sense of control, and confidence that the system knows your location accurately, are important. Latency in positioning affected this, as well. Figure 4.01 illustrates the relationships between user perceptions and key elements of the SpaceBook system that affected them.

On exploration tasks, the Visibility Engine was anticipated to be important, but its presence seemed to make no difference to users' satisfaction with the amount of information they could access, largely because input processing failure overrode Visibility Engine performance.

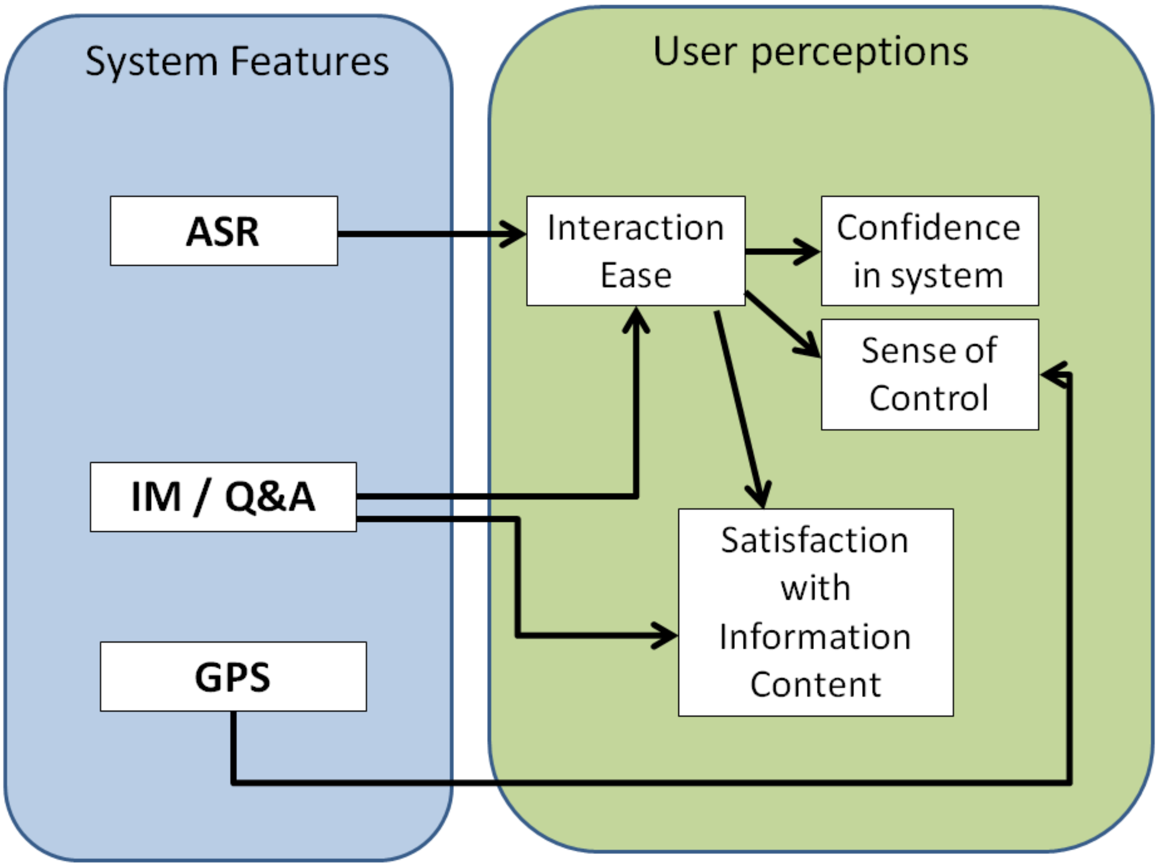


Fig. 4.01. Relationships between user perceptions and critical elements within the Spacebook System that affect them.

Automatic Speech Recognition and Speech Understanding

The high failure rates in the ASR were due to a number of causes. Firstly, it was 'always listening', and this, combined with the ambient conditions in the test location (very windy, with high tourist footfall, and winter-scheduled roadworks), meant that the system picked up a lot of false positive input, or genuine user-input was misheard. A wide range of accents was represented in the user cohort, which may also have led to high ASR failure rates, and the language understanding module did not use machine learning techniques, so the iterative 'hand made' parser was somewhat limiting. The quality of ASR from the first prototype was upgraded based on the utterances collected in the first evaluation, but this did not account for the full range of user utterances encountered in the final evaluation, which took place in a different area of the city and used different tasks. Consequently, the poor success rate of the input processing occluded, to a certain degree, the capacity of the designed evaluation to differentiate the functionality of the three systems, from the user's perspective.

Interaction and Q&A

Cambridge University's parsing element of the system was not incorporated, so the quality of co-referencing within the system meant that there was some confusion between what the person was asking for, and the information they received (see Usability). The psycholinguistic element of the interaction between user and system also affected information retrieval, particularly when there were differences between the way in which the model worked, and the user's expectations of what SpaceBook could actually understand. Users expressed frustration with a system that they perceived as being obtuse, but which was, in fact, merely being literal. As is well understood (Le Bigot *et al.* 2007; Branigan *et al.* 2011) when a user begins to doubt the capacity of a system, they tend to revert to simplistic forms of query, by using simpler syntax, repeating instructions and hyper-articulating. This was observed in a number of instances in this evaluation where initial input processing problems led to mounting frustration being exhibited by the user, as they attempted to make themselves understood. Response to Q&A was rather general in subject with a tendency not to satisfy a deeper curiosity of the subject. There was an absence of depth and detail to the responses, with users feeling that the system was frequently telling them the obvious or already known (description, or name of a point of interest). In some instances, a large city vista was presented to the user, but SpaceBook was inadequate in describing that vista. Because of the close coupling of the Q&A with the visibility engine, when the visibility engine was turned off, the Q&A became very poor. A better solution would be to link Q&A with a simple proximity-based approach.

Latency in the system, combined with the challenge of determining when to push exploratory information whilst prioritising wayfinding instructions led to various moments of confusion and frustration. It was always the intention to combine navigation with Q&A, but this proved to be more challenging than anticipated. The balance between push and pull is difficult to judge. Generally, there are many different ways by which subjects referred to objects of interest. Trying to develop a system that might cater to all these different forms was not easy.

SpaceBook was intentionally dialogue-only and this led to specific design challenges, such as determining the amount of redundancy required in the giving of instructions. Positioning latency causes users to seek reassurance that the system 'knows' where they are, and some users expressed a preference for additional forms of visual reference and interaction to provide this, and sustain confidence in the system. This preference for mixed media was somewhat driven by the high ASR failure rate: if users felt that they could not get the system to understand *them*, they were less likely to trust that they could get it to take them where they needed to be - they felt as though their control was slipping

Whilst many technical problems were overcome (fast retrieval from a rich database, pedestrian tracking, rich descriptions of points of interest in relation to topological modelling), psycholinguistic issues would be deserving of further development in the future. For example, SpaceBook did not differentiate between different types of users, their knowledge, gender, age, familiarity with the space or preferences (in terms of form and detail of description, types of landmarks pushed, preferred route *etc.*). However, incorporation of these options through either pre-task manual settings, or *en route* machine learning personalisation may push the system out of its current 'hands free, eyes free' mode of operation and thus reduce the distinction between SpaceBook and alternative existing smartphone applications.

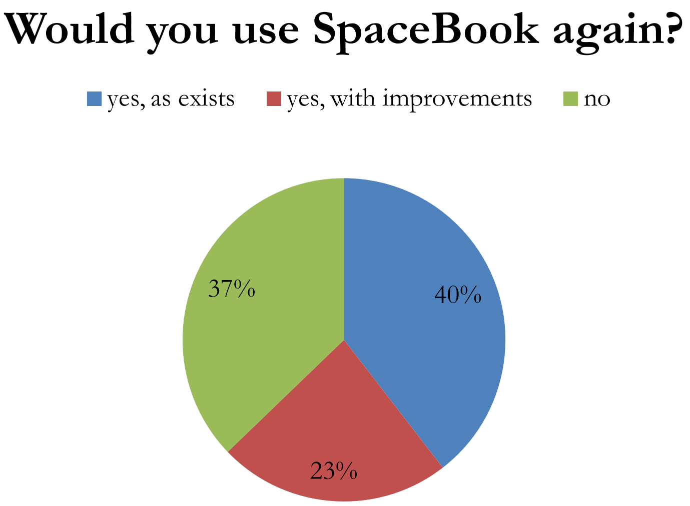


Fig. 4.02. User responses when asked, at the end of the experiment, if they would use the SpaceBook system again

There has been a growing recognition over the past decade of the importance of user-centered evaluations of information retrieval systems (Spink 2002; Diaz *et al.* 2008), so this report finishes with the final assessment from the users who tested the Spacebook system in this evaluation. Despite the problems outlined above, when asked nearly two thirds of users said that they would use SpaceBook again, either in its existing form, or with certain improvements (fig. 4.02), and only around one third of users said that they would not.

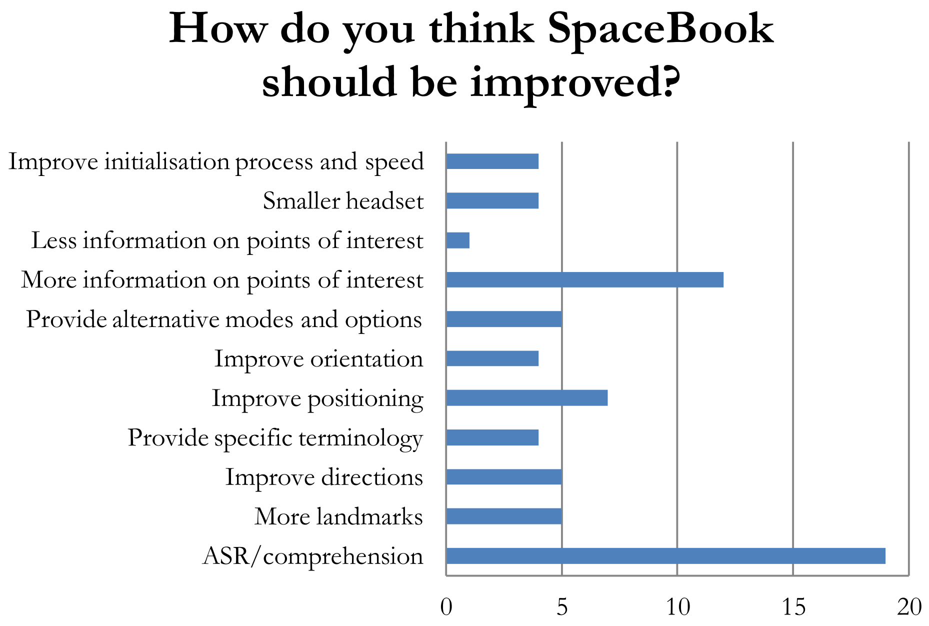


Fig. 4.03. User responses to the post-experimental open question, *How do you think SpaceBook should be improved? x*-axis = number of users.

When asked what improvements they felt should be made to SpaceBook, a variety of suggestions were put forward, with the two most important changes needing made being improvement in the input processing modules, and improvements in the amount of information that the system accesses and delivers to the user (fig. 4.03). This underlines the importance of the ease with which the user interacts with the system, and the expectations of information content, and how these affect overall impression. Users seem to be much more critical about these aspects of SpaceBook, and generally happier with the location and direction-finding capabilities of the system.

.

REFERENCES

Aman, F.; Vacher, M.; Rossato, S.; Portet, F., "Speech recognition of aged voice in the AAL context: Detection of distress sentences," *Speech Technology and Human - Computer Dialogue (SpeD), 2013 7th Conference on* , vol., no., pp.1,8, 16-19 Oct. 2013
doi: 10.1109/SpeD.2013.6682669

Bartie, P., Clementini, E. & Reitsma, F. (2013) A qualitative model for describing the arrangement of visible cityscape objects from an egocentric veiwpoint. *Computers, Environment and Urban Systems* **38**: 21-34

Branigan, H., Pickering, M. J., Pearson, J., & McLean, J. F. (2011) Linguistic alignment between people and computers. *Journal of Pragmatics* **42**(9): pp. 2355-2368

Brewster, S. (2002) Visualisation tools for blind people using multiple modalities. *Disability and Rehabilitation* **24**(11-12): 613-621

Diaz, A., Garcia, A. & Gervais, P. (2008). User-centred versus system-centred evaluation of a personalization system. *Information Processing and Management.* **44**: pp.1293-1307

Harvey Sacks, Emanuel A. Schegloff and Gail Jefferson "A Simplest Systematics for the Organization of Turn-Taking for Conversation" *Language*, Vol. 50, No. 4, Part 1 (Dec., 1974), pp. 696-735. Published by: Linguistic Society of America. Article DOI: 10.2307/412243.

le Bigot, L., Rouet, J., Jamet, E. (2007) Effects of speech- and text-based interaction modes in natural language human-computer dialogue. *Human Factors* **49**(6): pp. 1045-1053

M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens, Automatic speech recognition and speech variability: A review, *Speech Communication*, Volume 49, Issues 10–11, October–November 2007, Pages 763-786, ISSN 0167-6393, http://dx.doi.org/10.1016/j.specom.2007.02.006.

Montello, D. R. (1997) The perception and cognition of environmental distance: direct sources of information. Ed. Hirtle, S. C. & Frank, A. U. Conference: *International Conference on Spatial Information Theory - A Theoretical basis for GIS location.* Pennsylvania Oct. 15-18 1997. Spatial Information Theory: A Theoretical Basis for GIS Book Series. *Lecture Notes in Computer Science* **1329**: 297-311

Spink, A. (2002) A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing and Management* **38**: 401-426

Weiser, M. & Brown, J. S. (1995). *Designing Calm Technology*. Xerox PARC.

Yang, K., Evens, M. & Trace, D. A. (2008) Using a Java Dynamic Tree to manage the terminology in a suite of medical applications. *Methods of Information in Medicine* **47**(6): 499-504

APPENDICES

APPENDIX 1: METHODOLOGY - TASK ROUTES AND QUESTIONNAIRES

A1.1 *EN ROUTE* NAVIGATION TASK QUESTIONNAIRE

a) “I completed the task.” Yes/ No/ Don’t know.

The following questions answered using the Leikert scale, below

1	2	3	4	5	6	7

Strongly
disagree

Neither
agree nor
disagree

Strongly
agree

b) " I found the task easy."

c) "I felt in control of my journey."

d) "I felt confident I was going to reach my destination."

e) Can you mark where on this map [map provided to subject] you felt you knew where you were going?

f) "I found SpaceBook easy to understand."

g) " The speech timing was..."

1	2	3	4	5	6	7

Much too slow

Just right

Much too fast

h) "I found what SpaceBook said interesting."

i) "The landmarks SpaceBook used to direct me were obvious and useful."

j) "SpaceBook understood me."

k) "I felt as if SpaceBook was responding to my questions."

l) "I was always sure what SpaceBook was talking about."

A1.2 *EN ROUTE* DISCOVERY TASK QUESTIONNAIRE

a) What did you find out? [open response format]

b) SpaceBook gave me the right amount of information

1	2	3	4	5	6	7
Strongly disagree			Neither agree nor disagree			Strongly agree

c) Could you see things you would have liked to know about that SpaceBook didn't talk about? [open response format]

d) Tell me an interesting bit of information you just learned. [open response format]

A1.3 POST-EXPERIMENTAL DEBRIEFING QUESTIONNAIRE

Of the following three navigation targets: Camera Obscura, Black Watch Monument, St Giles Cathedral...

a) Could you tell me which place you found easiest to find? and Why?

b) On which of these tasks did you find SpaceBook most helpful?

c) Was there anywhere you felt it was particularly unhelpful?

d) Where/ On which task did you feel most in control? and Why?

e) Did you feel Spacebook influenced that feeling of control one way or the other?

f) With which task did you feel most confident you were going to reach you destination?

g) How much do you think SpaceBook contributed to that confidence?

h) For which task did you find Spacebook easiest to understand?

i) For which task did you find what Spacebook said most interesting?

j) Could you tell me which landmarks, as chosen by Spacebook to direct you, you found the most helpful?

k) Did you feel there were times when Spacebook understood you well?

l) Did you feel there were times when it *didn't* understand you?

m) For which task did you most enjoy using SpaceBook and why?

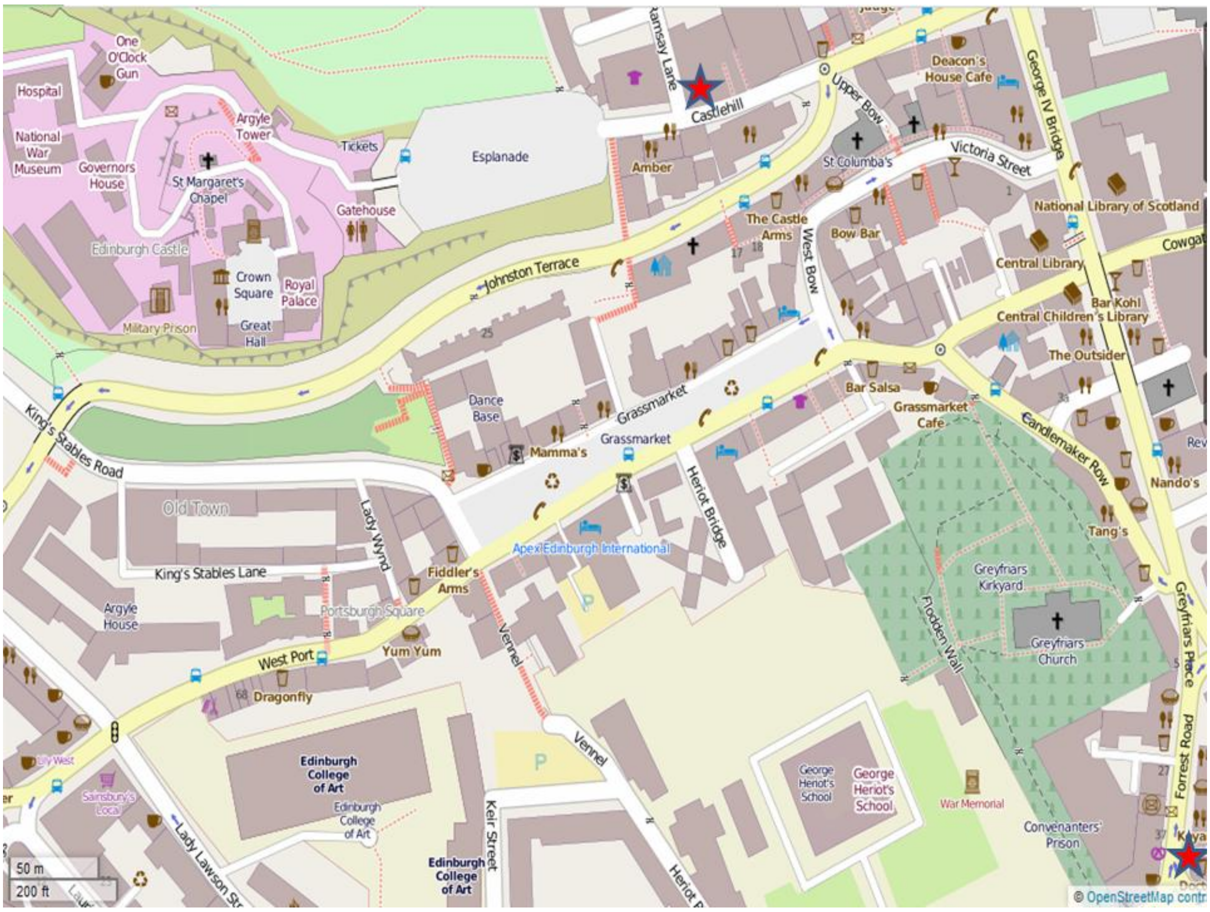
n) On which task did you least enjoy using Spacebook, and why?

o) Would you use SpaceBook again?

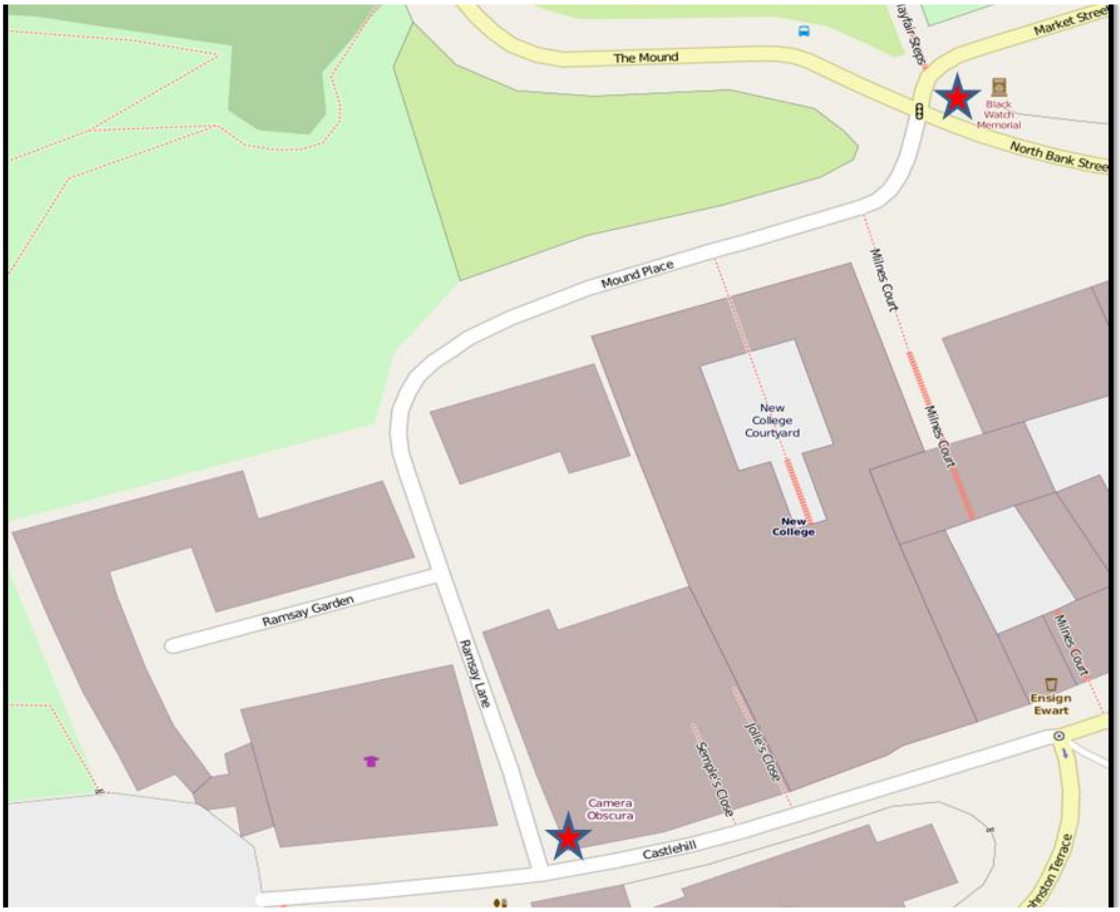
p) How do you think SpaceBook should be improved?

A1.4 NAVIGATION TASK MAPS

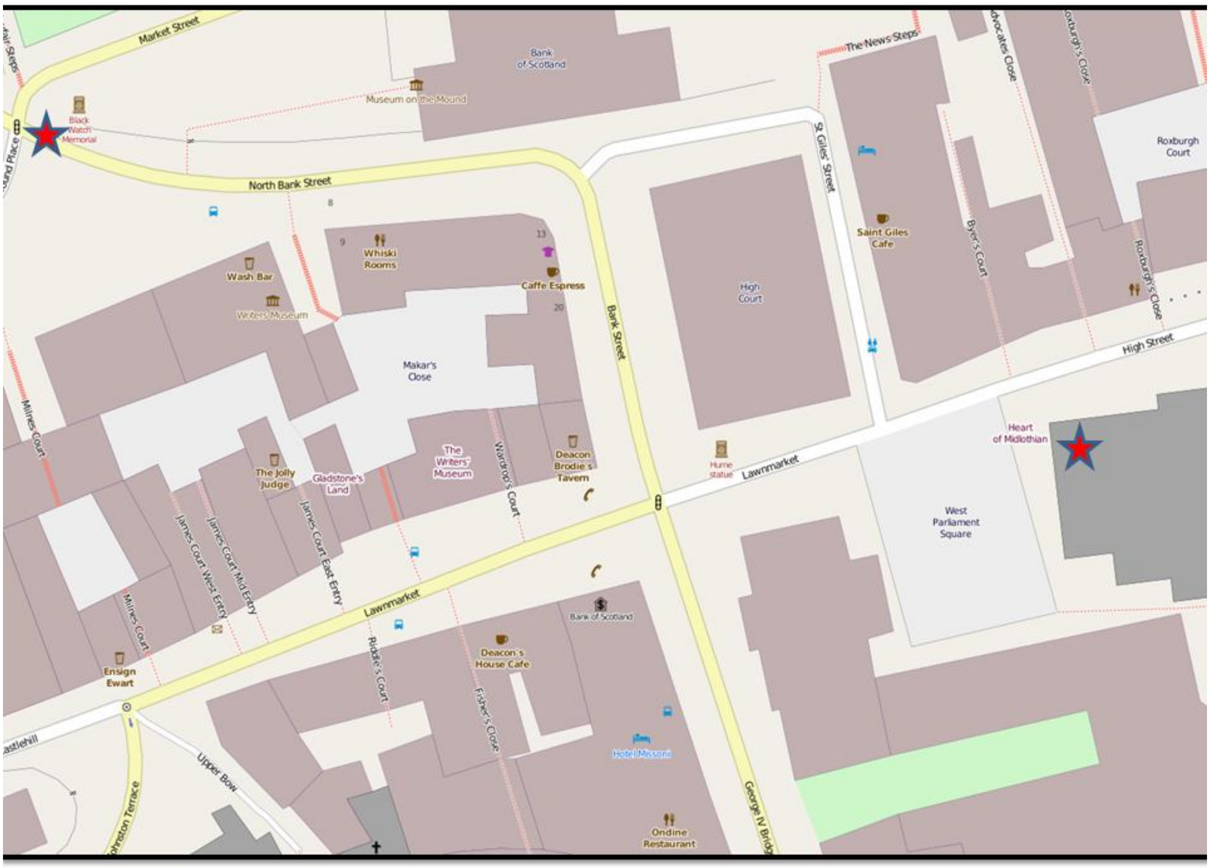
1.4a) Area map for Navigation Task 1. Red stars indicate start position (Doctor's Pub, on Forrest Road) and target (Camera Obscura on Castle Hill)



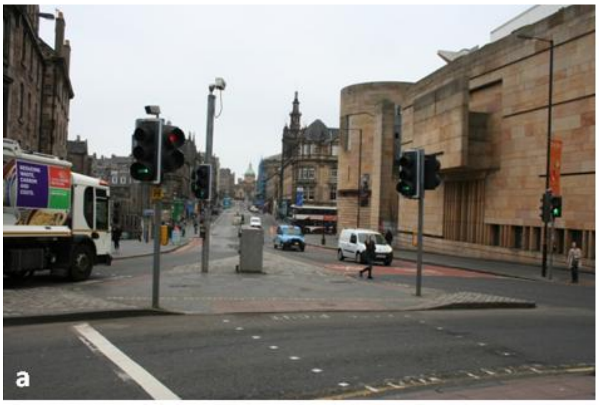
1.4b) Area map for Navigation Task 2. Red stars indicate start position (Camera Obscura, on Castle Hill) and target (Black Watch Monument on Market Street)

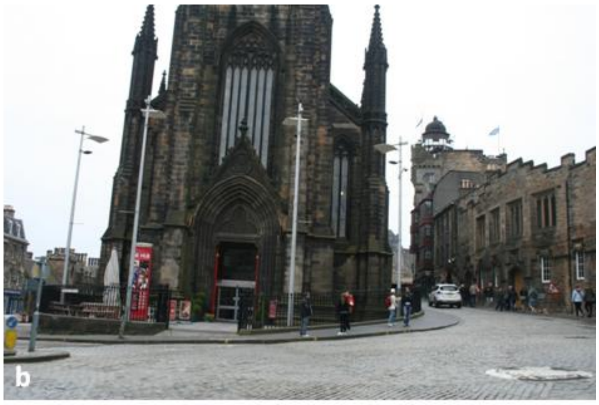


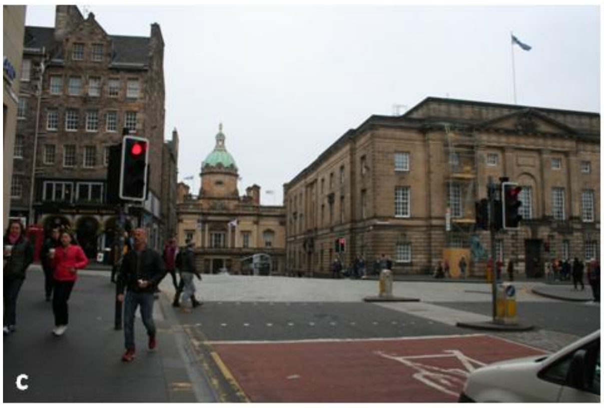
1.4c) Area map for Navigation Task 3. Red stars indicate start position (Black Watch Monument on Market Street), and target (St Giles Cathedral, in West Parliament Square)

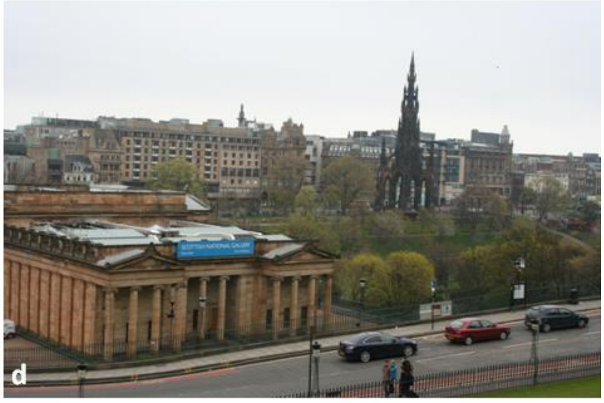


APPENDIX 2: IMAGES OF LANDMARKS, VISTAS AND OBJECTS OF INTEREST WITHIN THE USER TASKS









Key to images:

a) Navigation task 1; junction on Forrest Road
b) Navigation task 1: junction on Castle Hill, showing the Hub landmark (foreground) and the Camera Obscura target at the back right of the image
c) Navigation task 1: junction between George IV Bridge and the Royal Mile, showing the Bank of Scotland landmark in the centre background (green dome) and the High Court of Justiciary on the right
d) Navigation task 2: city vista from Ramsay Lane, looking North East, showing the National Gallery of Scotland in the foreground, Princes Street in the background, and Princes Street Gardens and the Scott Monument in the middle right of the image.
e) as a), but from a different angle
f) as b), but from a different angle
g) Milne's Court and j) Roxburghe Place: 2 examples of an Edinburgh 'Close'
h) Navigation task 2: Black Watch Monument target
i) Navigation task 3: showing West Parliament Square, with the High Court of Justiciary and the statue of David Hume on the left side of the image, and St Giles' Cathedral (target) in the centre
j) see g)
k) Explore task; showing Adam Smith statue in foreground and City Chambers in the background
l) Explore task: Tollbooth by St Giles' Cathedral

e
Edinburgh
f
n Shop
MILNES COURT
WEE SCOTLAN
g
h

i

BAGPIPES
ANGELS with BAGPIPES
j

ADAM SMITH
k

l

APPENDIX 3: RESULTS TABLES

A3.1 ITEM RESPONSE FREQUENCY DISTRIBUTIONS

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
I found the task easy	.208	30	.002	.918	30	.024
I felt in control of my journey	.177	30	.017	.923	30	.032
I felt confident I was going to reach my destination	.229	30	.000	.841	30	.000
I found SB easy to understand	.278	30	.000	.817	30	.000
The speech timing was...	.388	30	.000	.689	30	.000
I found what SB said interesting	.152	30	.075	.943	30	.109
The landmarks SB used were obvious and useful	.170	30	.027	.906	30	.012
SB understood me	.143	30	.119	.918	30	.024
I felt as if SB was responding to my questions	.180	30	.014	.880	30	.003
I was always sure what SB was talking about	.193	30	.006	.824	30	.000
Explore b	.288	30	.000	.841	30	.000
Vista b	.156	30	.060	.923	30	.032
I felt as if SB was responding to my questions M	.231	30	.000	.855	30	.001
I was always sure what SB was talking about M	.202	30	.003	.801	30	.000

a. Lilliefors Significance Correction

A3.2 ITEM RESPONSE - SYSTEM ASSOCIATION ANALYSIS

Item	Chi value	df	Cochran's Q	df	P	N
I completed the task	0.6077	2	-	-	-	-
I found the task easy	1.114	2	1.130	2	0.568	42
I felt in control of my journey	5.054	2	6.320	2	0.042*	42
I felt confident I was going to reach my destination	1.177	2	1.300	2	0.522	42
I found SB easy to understand	2.825	2	2.800	2	0.247	42
The speech timing was…	1.927	2	1.077	2	0.584	42
I found what SB said interesting	1.495	2	1.727	2	0.422	42
The landmarks SB used to direct me were obvious and useful	0.343	2	0.231	2	0.891	42
SB understood me	0.343	2	0.316	2	0.854	40
I felt as if SB was responding to my questions	0.483	2	2.480	2	0.289	36
I was always sure what SB was talking about.	1.002	2	1.520	2	0.468	41
Explore – SB gave me the right amount of information	7.491**	1	-	-	-	-
Vista – SB gave me the right amount of information	8.673**	1	-	-	-	-

* sig at p=0.05; ** sig at p=0.01

A3.3 SYSTEM AND AGE EFFECTS ON ITEM REPONSE (GLM ORDINAL REGRESSION ANALYSIS RESULTS)

Item	**Test of parallel lines** **(χ^2 $_{[df]}$)**	**Model fit**			**Goodness of fit:**			**Nagelkerke's pseudo R^2**
		χ^2 statistic	**df**	**Sig.**	**Pearson χ^2 statistic** **[Deviance χ^2]**	df	sig	
I completed the task		0.613	3	0.893	3.133 [4.205]	2	0.209 [0.122]	0.10
I found the task easy	$20.68_{[15]}$	2.299	3	0.513	25.268 [29.718]	27	0.559 [0.327]	0.018
I felt in control of my journey	$17.103_{[15]}$	5.181	3	0.159	23.223 [26.446]	27	0.673 [0.494]	0.041
I felt confident I was going to reach my destination	$16.830_{[15]}$	3.971	3	0.265	26.618 [28.917]	27	0.485 [0.365]	0.031
I found SB easy to understand	$5.188_{[15]}$	2.161	3	0.540	26.805 [27.023]	27	0.474 [0.463]	0.017
The speech timing was…	$26.357_{[12]}$*	4.636	3	0.200	25.984 [24.948]	22	0.252 [0.300]	0.039
I found what SB said interesting	$16.679_{[15]}$	2.974	3	0.396	24.177 [26.539]	27	0.620 [0.489]	0.024
The landmarks SB used to direct me were obvious and useful	$14.131_{[15]}$	1.383	3	0.710	30.077 [34.888]	27	0.311 [0.142]	0.011
SB understood me	$18.231_{[15]}$	2.593	3	0.459	24.954 [27.137]	27	0.577 [0.456]	0.021
I felt as if SB was responding to my questions	$8.580_{[12]}$	2.770	3	0.429	23.995 [30.428]	22	0.347 [0.108]	0.024
I was always sure what SB was talking about.	$7.166_{[12]}$	0.968	3	0.809	16.519	22	0.789	0.008

					[18.137]		[0.698]	
Explore – SB gave me the right amount of information	36.311$_{[10]}$**	14.517	2	0.001	9.825 [10.722]	16	0.876 [0.826]	0.359
Vista – SB gave me the right amount of information	12.867$_{[10]}$	7.497	2	0.024	23.735 [25.964]	16	0.095 [0.055]	0.192

A3.4 SYSTEM AND AGE EFFECTS ON SCALE RESPONSE (REPEATED MEASURES ANOVA RESULTS)

	Test for sphericity		**Residual normality**	**System**		**Age**		**System*Age**	
Scale	**Mauchley's W$_{[df]}$**	**P**		**F$_{[df]}$**	**P**	**F$_{[df]}$**	**P**	**F$_{[df]}$**	**P**
Confidence and Control	0.981$_{[2]}$	0.679	y	1.304$_{[2,40]}$	0.283	1.174$_{[1,41]}$	0.285	0.311$_{[2,40]}$	0.734
Information Content	0.948$_{[2]}$	0.341	y	0.365$_{[2,40]}$	0.697	0.484$_{[1,41]}$	0.491	1.776$_{[2,40]}$	0.182
Interaction Ease	0.917$_{[2]}$	0.176	y	0.479$_{[2,40]}$	0.623	0.703$_{[1,41]}$	0.407	0.086$_{[2,40]}$	0.917
Interaction Ease (modified)	0.947$_{[2]}$	0.336	y	1.078$_{[2,40]}$	0.350	0.440$_{[1,41]}$	0.511	0.714$_{[2,40]}$	0.496
Information Content (including Explore and Vista responses)	-	-	y	0.00$_{[1,34]}$	0.989	6.495 $_{[1,34]}$	0.016*	0.390$_{[1,33]}$	0.537

Test results for repeated measures Analysis of Variance of scale response, with 'System' as within-subjects independent factor, and 'Age' as between-subjects independent factor

A3.5 TASK EFFECTS ON SCALE RESPONSE

Scale	Variation source	df	*F*	*p*
Confidence and Control	Age	1,110	2.336	0.129
	System	2,110	0.994	0.374
	Task	2,110	2.282	0.107
	Age * System	2,110	0.367	0.694
	Age * Task	2,110	0.045	0.955
	System * Task	4,110	0.264	0.901
	Age * System * Task	4,110	0.091	0.985
Interaction Ease	Age	1,110	2.018	0.158
	System	2,110	0.154	0.857
	Task	2,110	3.050	0.051
	Age * System	2,110	0.148	0.862
	Age * Task	2,110	0.227	0.797
	System * Task	4,110	0.321	0.863
	Age * System * Task	4,110	0.146	0.964
Information Content (2-item scale)	Age	1,110	1.468	0.228
	System	2,110	0.504	0.606
	Task	2,110	2.964	0.056
	Age * System	2,110	0.769	0.465
	Age * Task	2,110	0.609	0.545
	System * Task	4,110	1.271	0.286
	Age * System * Task	4,110	1.367	0.250
Information Content (4-item scale)	Age	1,34	6.495	0.016
	System	1,34	0.000	0.989
	Age * System	1,34	0.390	0.537