Experimental evaluation of mobile interactions

Stephen Brewster

Glasgow Interactive Systems Group
School of Computing Science
University of Glasgow

stephen@dcs.gla.ac.uk

Overview

Issues with standard usability assessments

Field experiments

- Design of experiments
- Participants and safety
- Measures

Case studies of mobile experiments

Traditional usability studies

Comparison to traditional evaluation

What do lab-based studies tell you about mobile use?

- Environment is very different: no distractions, quiet, no reflection on screen, weather, no dodging pedestrians or traffic
- Cognitive load is different
- Ecological Validity

Cannot completely control experimental environment

Cost (time/money) of running experiment is high

Comparison to ethnographic studies

Field experiments can give quantitative results

Objective comparison of different devices or interfaces

Compare performance (times/errors)

Ethnographic studies give you ‘bigger picture’

Qualitative data may be important for real use

Longer term effects might not show up in shorter term experiments

New applications and services

- Location based services
  - Rely on user's location to trigger an application or service
  - How do you test these in the lab?
  - Real effects of GPS errors
- Mobile tourist guides
  - How do you evaluate it without giving it to tourists to try out around the city?
- In-car / pedestrian navigation systems
  - Designed to be used with users on the move

Field experiments

Draw together the advantages of lab-based studies with the validity of real mobile use

Within-subjects designs if possible

- Saves time and increases validity of results

Be careful of learning effects

- Cannot use the same route twice in a navigation interface
- Ensure routes are of same difficulty
- Counterbalancing important

Do lab study as well as real world one

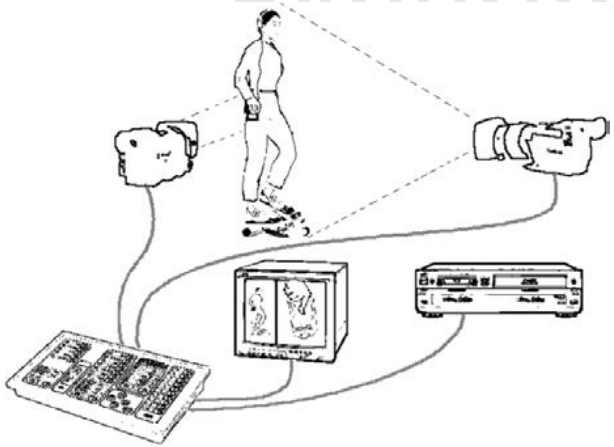
- See what affects mobility has on your interface

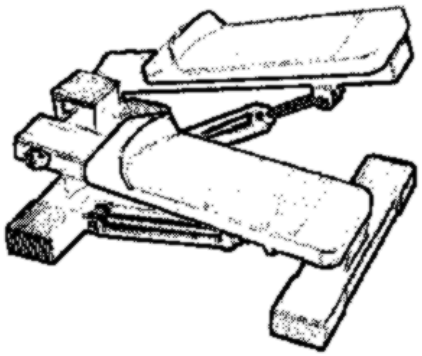
Experimental designs

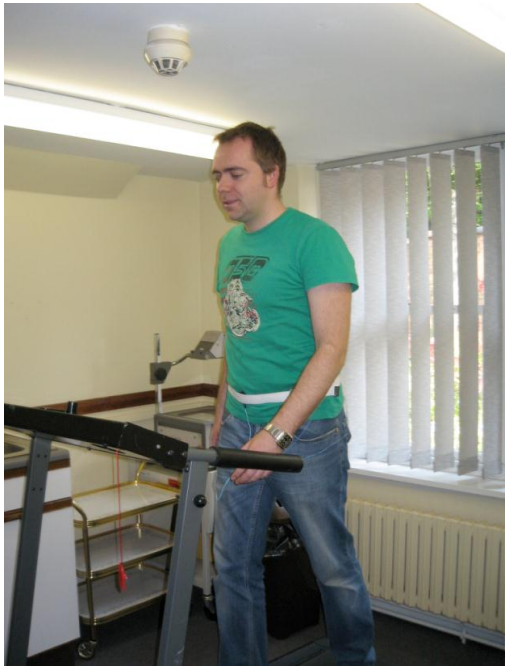
Different levels of mobility

- Treadmill: physical effects of movement, but not cognitive ones
  - Good if early prototypes not ready for mobile use
- Controlled walking routes: physical and cognitive effects of walking, safe, but not effects of ‘real’ environment
  - Good for keeping more control
- Field experiments: all the affects of the real world. Most realistic but hardest to control
  - Good for seeing real world issues

Experimental designs







Controlling variables

Need to control things enough to collect valid data

- Tricky in the lab, hard in the real world

Unusual occurrence in one run of experiment

- Messes up results

Try to ensure such affects are spread across whole experiment to even things out

- Spread the variation out

Can control variations to some extent

- Test at same times of day, avoid really busy times, …

Participants

May need more participants than a lab study to deal with variations introduced by the environment

- Loss of control of experimental environment

Can be harder to recruit them as they have to work harder

- Payment!

Can be tiring if they have to walk a long way

- Different user groups

Must be very careful with your participants

- Ethics approval / informed consent
- Lots of pilot testing to ensure everything works

Safety

- Participants are moving, walking, driving, running,…
  - Need to ensure their safety: ethics approval
- Avoid as many hazards as possible
  - Routes that avoid steep steps
- For outdoor studies use a park rather than a street
  - Less change of getting run over
- Experimenter can walk behind participant to ensure safety
  - Be prepared to stop expt if a problem/danger occurs

Choice of test environment

- Depends on what you want to know
- Chose environment based on use of actual interface
  - Will use the interface walking or sitting? Indoors or outdoors?
- May be limited by technology
  - Using GPS? Test outdoors
  - Lots of wires? Controlled indoor walking
- May need permission for testing
  - Private property
  - Subway is owned by SPT

Measures

- Dependent variables
  - You manipulate Independent Variable(s) and look for effects in Dependent Variable(s)
  - Record as much as possible as it can be hard to know where effects will occur
  - Triangulation
- Times and Error rates
  - ‘Classic’ measures, good for mobile testing too
  - Easy to measure as can be recorded by computer
  - Could be interaction errors, navigation errors, …

Measures

- Note taking
  - Always important to take notes whilst experiment is going
  - Qualitative data
  - Record any unusual events that might affect results
- Post-test questionnaires
  - More qualitative data
  - Find out more about what people think, how they found the task, how they liked your interface
  - Always important to do these

Subjective workload

NASA Task Load Index (TLX)

“effort invested by the human operator into task performance” (Hart and Staveland)

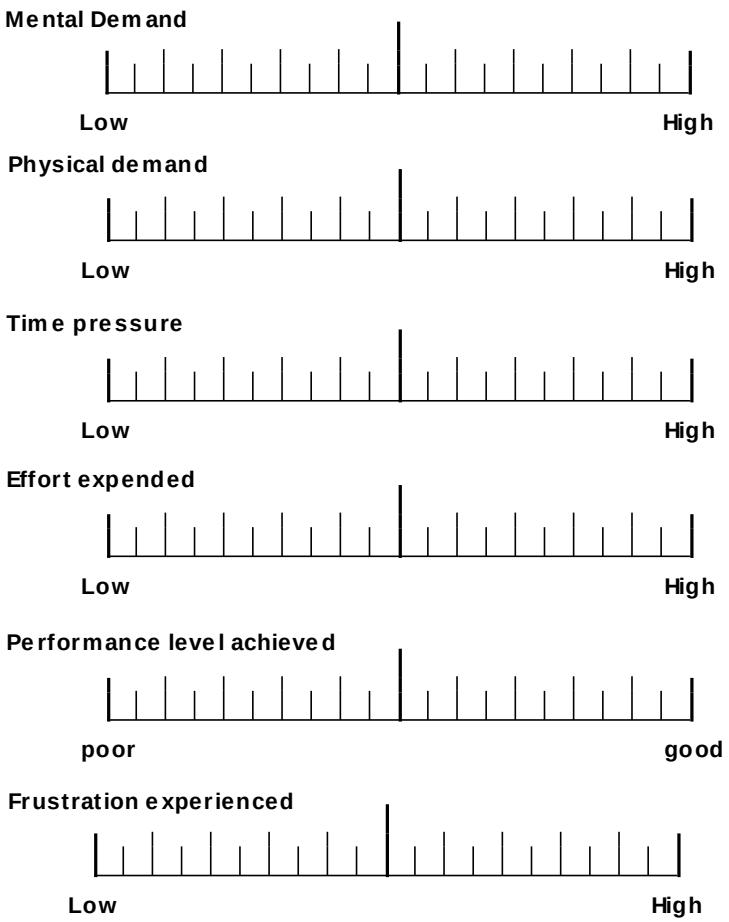
- Times and error rates may be good for one interface, but users may have to work much harder to use it
- Problems then come when used in real world
- An interface that lowers workload will be more successful in mobile use

Subjective workload

Set of 6 rating scales with standard definitions

WORKLOAD Rating Scale Definitions		
Title	**Endpoints**	**Description**
Mental Demand	*Low/High*	How much mental, visual and auditory activity was required? (e.g. thinking, deciding, calculating, listening, scanning, searching)
Physical Demand	*Low/High*	How much physical activity was required? (e.g. pressing, controlling)
Time Pressure	*Low/High*	How much time pressure did you feel because of the rate at which things occurred or the time limit imposed on the task? (e.g. slow, leisurely, rapid, frantic)
Effort Expended	*Low/High*	How hard did you work (mentally and physically) to accomplish your level of performance?
Performance Level Achieved	*Poor/Good*	How successful do you think you were in doing the task set by the experimenter? How satisfied were you with your performance?
Frustration Experienced	*Low/High*	How much frustration did you experience? (e.g. were you relaxed, content, stressed, irritated, discouraged)

Subjective workload



Percentage Preferred Walking Speed (PPWS)

Measures extent to which interaction disrupts walking

- The further users walk below their normal walking speed, the more negative the effect of the device
- Texting while walking vs phone call while walking

"… all pedestrians have a walking speed which they prefer ... The ability of any mobility aid to allow the pedestrian to walk at this preferred walking speed (PWS) is therefore argued to be a measure of its effectiveness". (Petrie)

PPWS

Get participants to walk route without using your interface

- Take average speed

Compare walking speed with and without your interface

- How much does it make them slow down?

Distance travelled/route taken

- Total distance walked/number of laps covered
  - Greater the distance walked the less the effect of the device (linked to PPWS)
  - How close to ‘ideal’ distance/route?
- Can help to identify location-based errors and points where difficulties arose
- How to measure
  - Pedometer (attached to belt), may need calibration
  - GPS (not inside though), gives actual route travelled

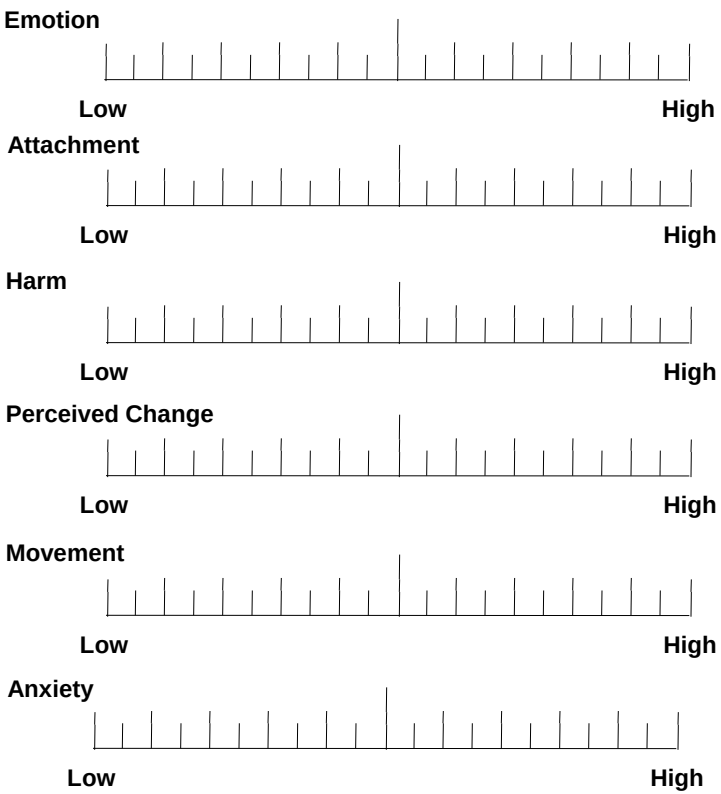
Comfort Rating Scale (CRS)

Measures participants' comfort wearing a device

- Developed to test wearable computers
- Use like TLX

Gives you some idea of how people feel when wearing or using new devices

CRS



Title	Endpoints	Description
Emotion	Low/High	I am worried about how I look when I wear this device. I feel tense or on edge because I am wearing the device.
Attachment	Low/High	I can feel the device on my body. I can feel the device moving.
Harm	Low/High	The device is causing me some harm. The device is painful to wear.
Perceived change	Low/High	Wearing the device makes me feel physically different. I feel strange wearing the device.
Movement	Low/High	The device affects the way I move. The device inhibits or restricts my movement.
Anxiety	Low/High	I do not feel secure wearing the device.

Comparing evaluation measures

Measure	What does it tell us?	How is it measured?
Timings	Performance	By the experimenter via a watch or stopwatch. Automatically by the device.
Errors	Performance. Particular sticking points in a task or design	By success in completing the tasks or answering questions correctly, Through experimenter observation, examining the route walked and the buttons pressed.
Perceived workload	Effort invested. User satisfaction	Through NASA TLX scales and other questionnaires and interviews.
Distance travelled and route taken	Depending on the application, these can be used to pinpoint errors and to indicate performance	Using a pedometer, GPS or other location-sensing system. By experimenter observation.
Percentage preferred walking speed (PPWS)	Performance	By dividing distance travelled by time to obtain walking speed, which is then compared with normal walking speed.
Comfort	User satisfaction. Device acceptability	Using the Comfort Rating Scale and other questionnaires and interviews.
User comments and preferences	User satisfaction and preferences. Particular sticking points in a task or design	Through questionnaires, interviews and think-alouds. Using open-ended questions and forced choice.
Experimenter observations	Many different aspects, depending on the experimenter and on the observations	Through observation and note-taking by the experimenter.

Case study: Gesture MP3 player

Designed TouchPlayer an MP3 player controlled by gestures

- Used non-visually

Gestures based on physical actions

- Sweep left -> right: next track
- Sweep right -> left: previous track
- Double tap: start/stop
- Circle clockwise: volume up
- Circle anticlockwise: volume down

Tested in lab and in controlled walking environment

Experiment

Two-condition within groups design

Interface type as independent variable

- Touchplayer vs standard PDA media player

15 participants

Dependent variables

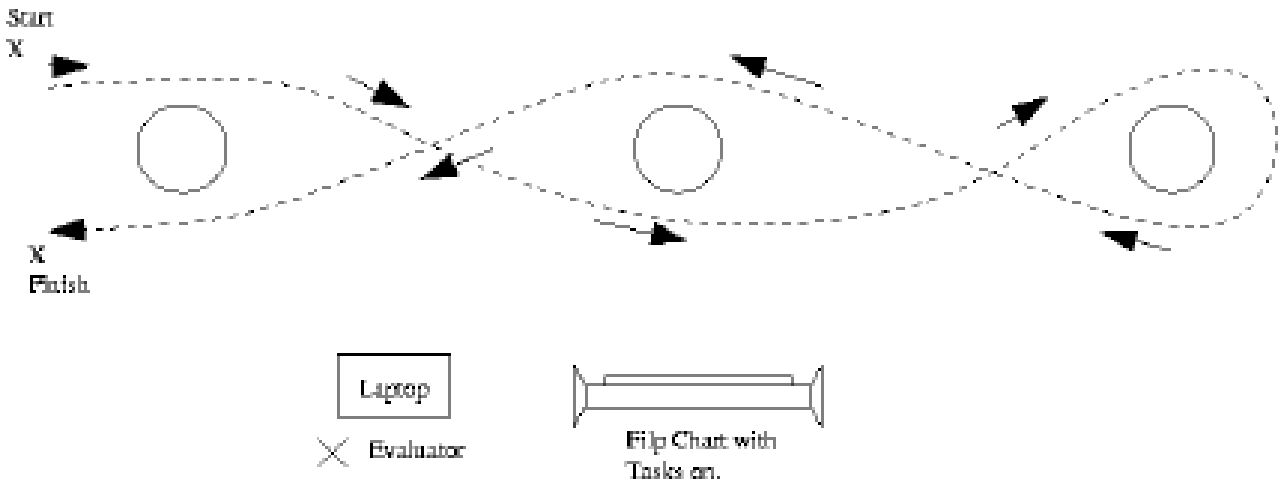
- Time to complete tasks, Error rates
- Subjective workload
- Percentage Preferred Walking Speed

Tasks

Perform a range of gestures on the device whilst on the move

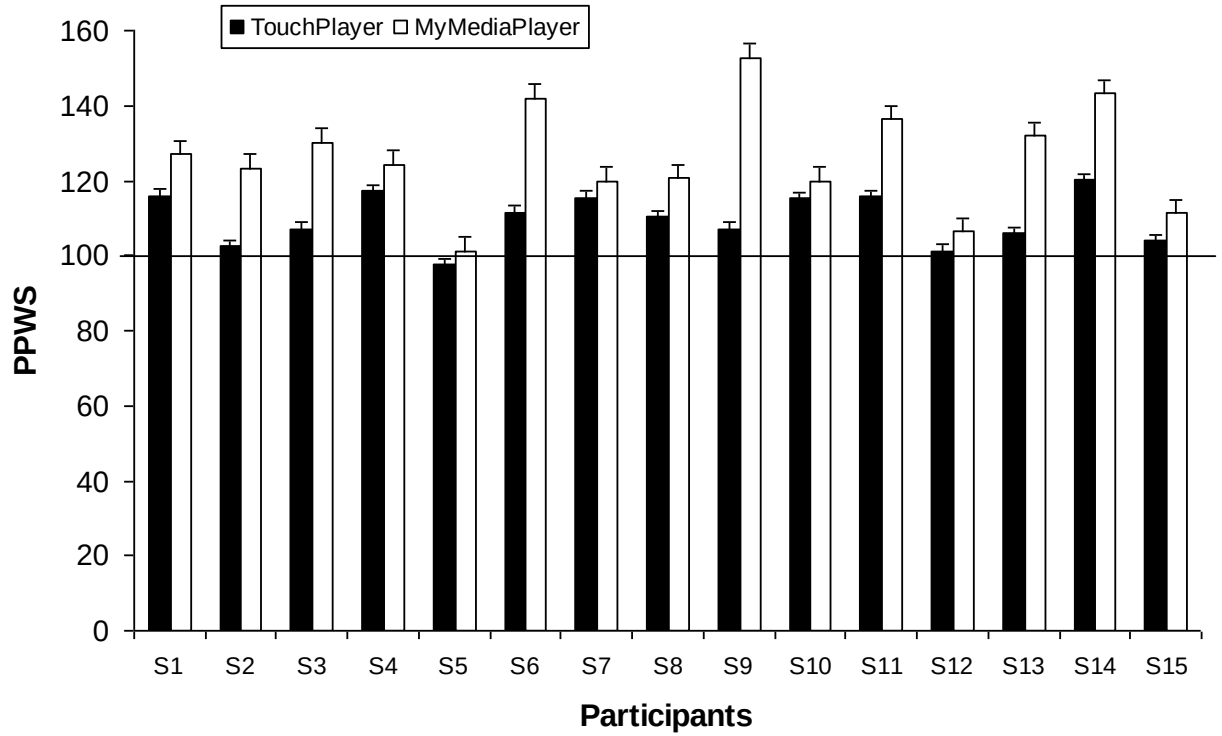
Controlled walking design

- Walk 17 x 8m laps of corridor avoiding obstacles
- Collected PWS at beginning



PPWS Results

PPWS significantly slower with standard Media Player



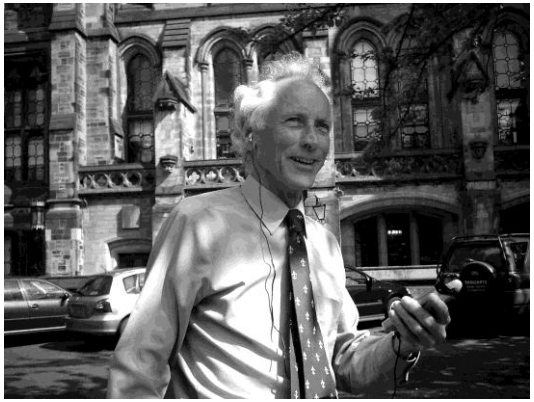
Case study: navigation aid for older adults

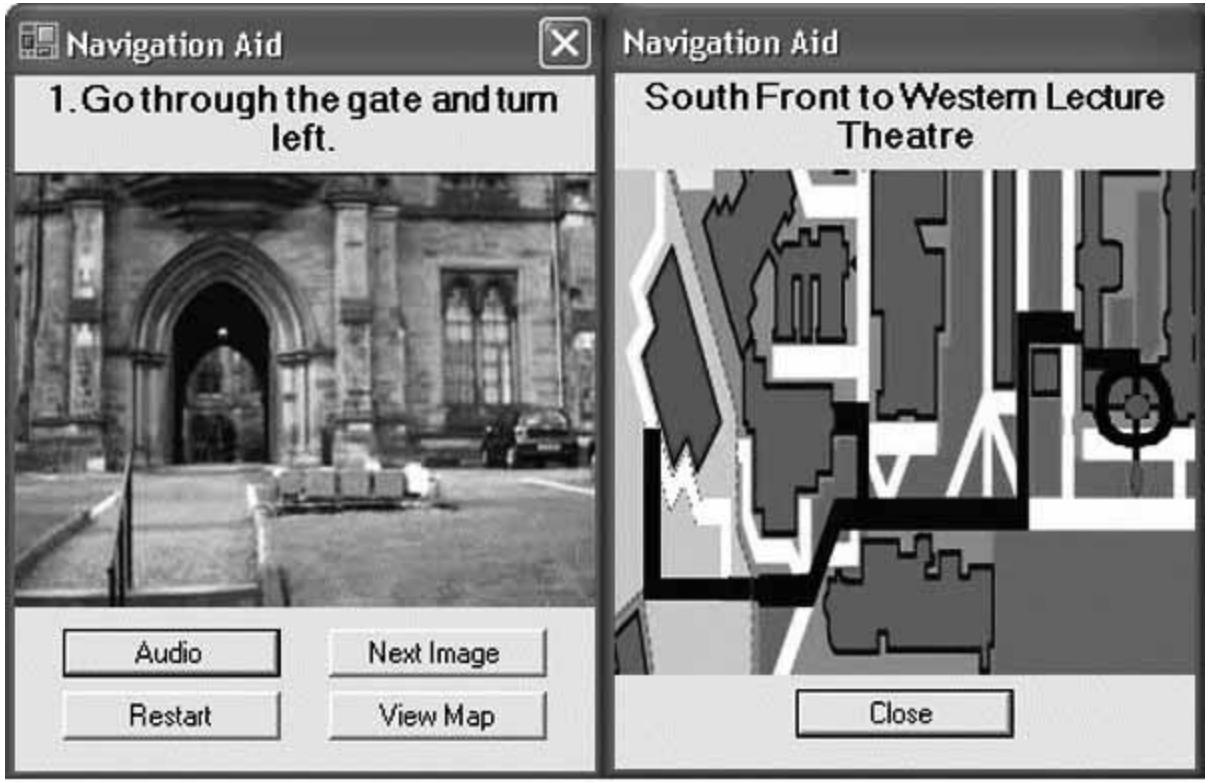
Designed a mobile navigation aid for older adults

Experiment compared different types of instructions and older vs younger users

- Photos, arrows, text, speech

Participants walked routes around the Uni (accompanied by experimenter)



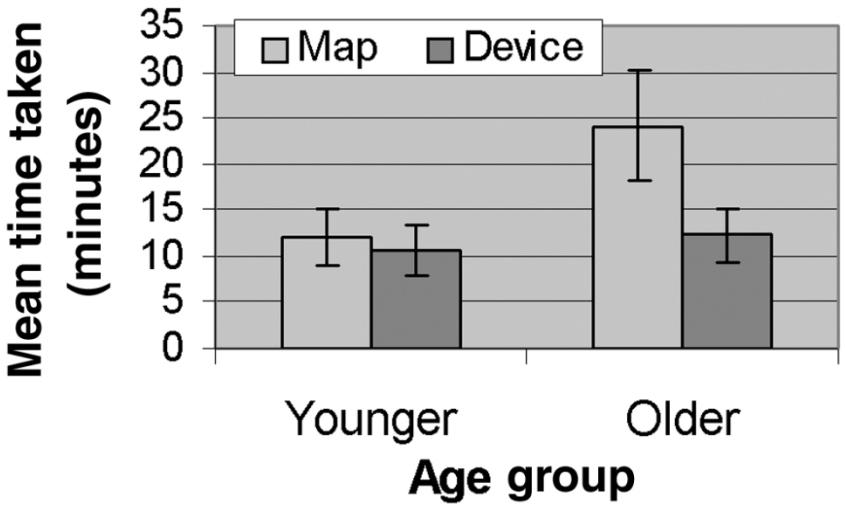

Navigation Aid
1. Go through the gate and turn left.
Audio
Next Image
Restart
View Map
Navigation Aid
South Front to Western Lecture Theatre
Close
(a)
(b)

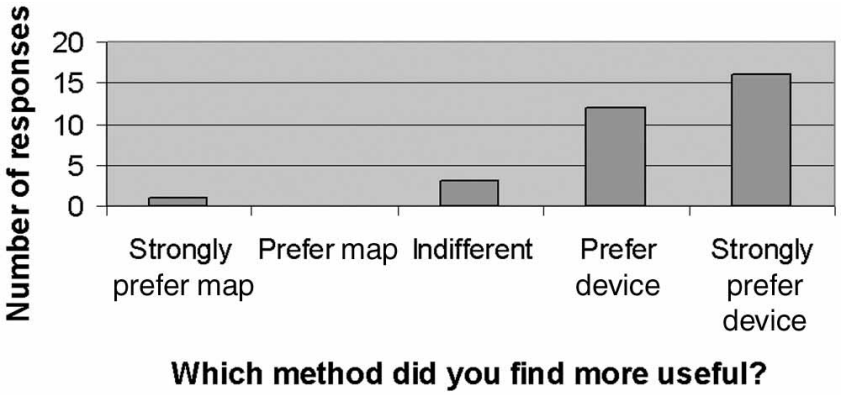
Navigation aid

With navigation aid older users as good as younger

Preference strongly for device

Had to change routes during expt due to building work





Summary

Traditional usability studies in the lab do not tell you enough about usage in the real world

Need experiments that bring in effects of real world

- Treadmill -> controlled walking -> real world studies

Be careful of your participants!

Many different measures you can use

- Collect as much as possible
- Sometimes hard to see where effects might be

Lab Next week

Don’t forget we are in the lab all of next week